RADARVL:FROM MODEL-GENERATED CANDIDATES TO EXPERT-VERIFIED SUPERVISION FOR DOMAIN SPECIALIZED VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Specializing general-purpose vision-language models (VLMs) to scientific visual tasks with high domain expertise requirements relies on large-scale, high-quality, and reliable domain supervision. However, constructing such supervision entirely through manual annotation by domain experts is costly, while directly using descriptions generated by strong VLMs as training labels may introduce erroneous supervision that is inconsistent with the original observations. To address this challenge, we focus on weather-radar understanding, an interdisciplinary task that requires professional meteorological knowledge, temporal visual reasoning, and multimodal interpretation, and propose an expert-verified synthetic supervision approach.

Based on this approach, we construct Radar-9K, containing 9,089 expert-verified convective events and 94,825 MRMS composite-reflectivity frames, and further specialize an 8B VLM into the weather-radar vision-language model RadarVL-8B. Experiments show that verified supervision substantially improves radar-domain understanding. On the MRMS test set, RadarVL-8B improves the score from 64.8 to 77.1. On unseen European OPERA radar observations, its expert score improves from 61.7 to 71.5 without additional fine-tuning. These results demonstrate that candidate annotations generated by strong VLMs can effectively support the specialization of weather-radar models, while reliability verification is critical for obtaining high-quality supervision and robust downstream performance.

1 INTRODUCTION

Weather radar imagery is a type of meteorological observation with strong domain-specific characteristics. Its accurate interpretation requires radar-meteorological knowledge and professional experience, and plays an important role in convective-system identification, evolution analysis, severe-weather monitoring, and nowcasting. Extensive radar research has established mature methods for storm identification and tracking (Dixon & Wiener, 1993; Heikenfeld et al., 2019; Feng et al., 2023) and precipitation nowcasting (Shi et al., 2015; Ravuri et al., 2021; Zhang et al., 2023). For continuous radar observations, a model must not only understand echo characteristics at individual observation times, but also analyze their structural changes and evolution over time.

In recent years, vision-language models (VLMs) have rapidly advanced in visual understanding and multimodal reasoning, providing a new pathway toward automated interpretation of weather radar observations. However, a clear domain gap remains when general-purpose VLMs are directly applied to weather radar, particularly in radar-specific semantics, continuous-event understanding, and observation-grounded analysis. Studies such as WeatherQA and RadarQA likewise show that multimodal models still face substantial challenges when reasoning over professional meteorological and radar observations (Ma et al., 2024; He et al., 2025).

Building a specialized vision-language model for weather radar requires large-scale, high-quality radar image–text supervision, while professional radar annotation depends heavily on domain expertise and is costly to construct entirely by hand. Using strong models to generate instructions or domain-specific descriptions has become an important strategy for reducing annotation cost and has been explored in general instruction tuning as well as biomedical, remote-sensing, and agricultural applications (Wang

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

et al., 2023; Li et al., 2023a; Kuckreja et al., 2024; Awais et al., 2025). This makes strong VLMs a natural source of candidate annotations that can subsequently be reviewed and corrected by domain experts.

However, model-generated outputs cannot be directly treated as reliable supervision. VLMs may still produce hallucinations or statements unsupported by the supplied observations (Li et al., 2023c; Guan et al., 2024; Jing et al., 2024), while unverified synthetic data may further propagate such errors to downstream models (Feng et al., 2025). The key challenge is therefore not only to scale radar supervision, but also to ensure that generated supervision remains consistent with the original radar observations and professionally reliable.

To address this issue, we propose an expert-verified synthetic supervision framework that treats radar analyses generated by a strong VLM as candidate annotations and improves their reliability through independent auditing, observation-grounded cross-verification, and domain-expert correction. Based on this framework, we construct the multimodal weather-radar dataset **Radar-9K** and train **RadarVL-8B**, a specialized vision-language model for professional understanding and analysis of continuous radar events.

The main contributions of this work are summarized as follows:

- We develop a specialized vision-language model for continuous weather-radar events, enabling the model to understand radar-echo characteristics and their temporal evolution from continuous observations while producing analyses grounded in observational evidence.
- We propose an expert-verified synthetic supervision method that transforms candidate annotations generated by a strong VLM into reliable radar-domain supervision through candidate generation, independent auditing, observation-grounded cross-verification, and expert correction.
- We construct **Radar-9K** and train **RadarVL-8B**, and evaluate their weather-radar understanding and cross-product generalization on MRMS and OPERA radar observations.

2 From VLM-Generated Radar Candidates to Reliable Supervision

Our framework progressively transforms continuous radar observations into expert-verified supervision and further uses the verified data for domain specialization through supervised fine-tuning. The overall pipeline, including event construction, candidate generation, independent auditing, cross-verification, and expert verification, is illustrated in Figure 1.

2.1 Scientific Input and Event Segmentation

We construct the radar supervision data from 2025 MRMS composite-reflectivity observations. Rather than treating radar observations at different times as independent samples, we organize observations belonging to the same tracked convective system into event-level sequences, so that echo displacement, structural changes, and evolution trends can be analyzed from multiple observation times.

To construct these events, convective cells are first identified from reflectivity fields using connected-component detection. Spatially nearby cells within the same observation time are then grouped into convective clusters. Clusters at successive observation times are associated through temporal tracking based on their spatial overlap, displacement, motion consistency, size, and reflectivity characteristics. Tracks satisfying the event-selection criteria are retained as convective events. After the primary tracks are established, lower-reflectivity echoes associated with the tracked system are used to extend the temporal and spatial context of the event, thereby retaining its developing and dissipating stages when supported by the observations.

Each resulting radar event is represented as a temporally ordered sequence

$$\mathcal{R}_i = \{I_i^1, I_i^2, \ldots, I_i^{T_i}\}, \tag{1}$$

where I_i^t denotes the radar observation at the t-th observation time and T_i is the number of observations retained for event i. This event-level representation preserves the correspondence of the same convective system across multiple observation times and serves as the multimodal input for subsequent supervision construction.

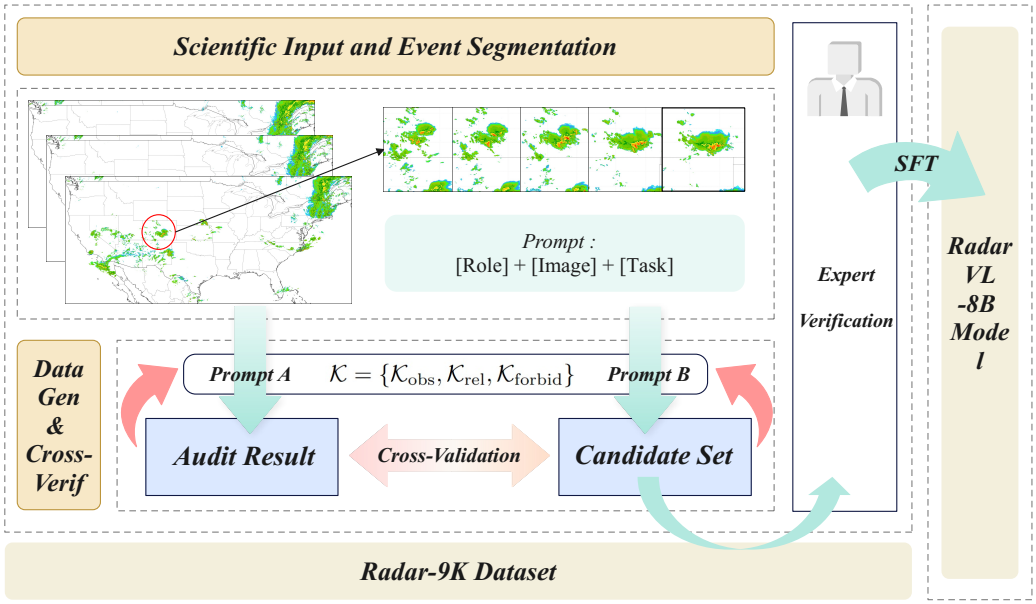


Figure 1: Overview of the expert-verified synthetic supervision framework for weather-radar VLM specialization. Radar observations are first organized into event-level sequences. A strong VLM then generates candidate annotations, while an independent auditor analyzes the same observations without access to the candidates. The two results are cross-verified under radar observation constraints, followed by domain-expert verification to construct Radar-9K. The verified supervision is finally used to specialize an 8B VLM through supervised fine-tuning, yielding RadarVL-8B.

Given $\mathcal{R}_i$, a strong VLM G_ϕ produces a candidate annotation

$$\hat{Y}_i = G_\phi\left(\mathcal{R}_i, q_i^{\text{gen}}\right). \tag{2}$$

A direct synthetic-supervision strategy would treat $(\mathcal{R}_i, \hat{Y}_i)$ as training data for supervised fine-tuning. In this work, however, $\hat{Y}_i$ is treated only as a candidate annotation rather than final supervision. We distinguish it from the verified annotation Y_i^* and decompose the candidate into atomic radar claims,

$$\hat{Y}_i = \{c_{i,r}\}_{r=1}^{L_i}, \tag{3}$$

where each $c_{i,r}$ describes a specific radar property or relation, such as echo intensity, morphology, organization, displacement, or temporal evolution. Subsequent reliability verification is therefore performed at the individual claim level.

2.2 Radar Observation Boundaries and Candidate Generation

Composite reflectivity supports some conclusions directly, some only through relations among frames, and does not support others. We encode these boundaries as

$$\mathcal{K} = \{\mathcal{K}_{\text{obs}}, \mathcal{K}_{\text{rel}}, \mathcal{K}_{\text{forbid}}\}. \tag{4}$$

$\mathcal{K}_{\text{obs}}$ covers directly observable properties including reflectivity intensity, echo location and extent, morphology, and organization. $\mathcal{K}_{\text{rel}}$ covers relations supported by the temporal sequence, including translation, strengthening, weakening, expansion, contraction, merging, splitting, and dissipation. $\mathcal{K}_{\text{forbid}}$ excludes conclusions that cannot be reliably established from composite reflectivity alone, such as radial velocity, vertical structure, thermodynamic causes, surface damage, or direct confirmation of hail and tornado occurrence.

Under these boundaries, Qwen3-VL-32B generates candidate claims together with evidence associations:

$$\hat{Z}_i = G_\phi(\mathcal{R}_i, q_i^{\text{gen}}; \mathcal{K}) = \{(c_{i,r}, e_{i,r})\}_{r=1}^{L_i}, \tag{5}$$

where $e_{i,r}$ identifies the frame, region, property, or temporal relation intended to support claim $c_{i,r}$. Importantly, these outputs are retained as candidates and are not yet accepted as training labels.

2.3 Independent Observation Audit

To provide an independent observational reference for evaluating the generator-produced candidate annotation, we introduce a separate audit process. An auditor A_ψ analyzes the same radar event directly from the original radar observations:

$$Z_i^A = A_\psi\left(\mathcal{R}_i, q_i^{\text{audit}}; \mathcal{K}\right), \tag{6}$$

where the auditor independently identifies the observed echo intensity, morphology, organization, displacement, and temporal evolution under the same radar observation boundaries defined by $\mathcal{K}$.

The audit is performed independently of the candidate-generation process. The auditor receives the original radar sequence and an audit-specific prompt, but is not shown the candidate annotation $\hat{Y}_i$, its structured claim–evidence representation $\hat{Z}_i$, or the generation history used to produce them. Consequently, Z_i^A provides an additional interpretation of the same radar observations that can be compared with the candidate claims during subsequent verification.

2.4 Cross-Verification and Expert Verification

The candidate annotation and independent audit are subsequently compared with each other and checked against the original radar sequence and the domain constraints $\mathcal{K}$. A verifier V_ω jointly examines the original radar sequence, the candidate annotation, the independent audit, and the radar observation boundaries:

$$\mathcal{E}_i = V_\omega\left(\mathcal{R}_i, \hat{Z}_i, Z_i^A, q_i^{\text{ver}}; \mathcal{K}\right), \tag{7}$$

where $\mathcal{E}_i$ denotes the inconsistencies or unsupported statements identified during cross-verification. Cross-verification is used to identify evidence–claim mismatches, contradictions between the two analyses, forbidden inferences, ordering errors when order is relevant, and missing evidence associations. These automatic procedures serve only as reliability-control mechanisms rather than as the final annotation authority. If a sample fails cross-verification, it is returned to the candidate generation and independent audit stage for regeneration and re-auditing until the verification requirements are satisfied.

Each sample that passes cross-verification is then reviewed against the original observations by a domain expert. The expert may accept, edit, delete, or rewrite individual radar claims, yielding the final expert-verified annotation Y_i^*. A subset of samples may additionally receive an independent review by a second domain expert to measure annotation reproducibility. When substantial disagreement occurs, a third expert may adjudicate the case based on the original radar observations. This design ensures full expert review of all accepted supervision while allowing inter-expert agreement to be independently assessed on a sampled subset.

2.5 Dataset Construction

The resulting Radar-9K contains 9,089 radar events and 94,825 radar frames, all derived from 2025 MRMS composite-reflectivity observations. We split the data at the event level into 6,362/1,363/1,364 train/validation/test events, respectively, ensuring that frames from the same tracked event never appear across different splits.

3 Experiments

3.1 Experimental Setup

Models and datasets. We use Qwen3-VL-8B as the base model for radar-domain specialization and Qwen3-VL-32B as the candidate-annotation generator. The final specialized model, RadarVL-8B, is obtained by supervised fine-tuning of Qwen3-VL-8B on the expert-verified Radar-9K supervision data.

The primary experiments are conducted on held-out MRMS convective events. To evaluate generalization beyond the radar product and geographic region used for specialization, we additionally construct an external test set consisting of 400 convective events from unseen European OPERA

composite-reflectivity observations (EUMETNET OPERA, 2024). No OPERA observations are used during supervision construction or fine-tuning.

For comparison with general-purpose VLMs of similar scale, we include InternVL3.5-8B (Wang et al., 2025), MiniCPM-V 4.5 (Yu et al., 2025), and GLM-4.1V-9B-Thinking (GLM-V Team et al., 2025). We additionally report direct inference with Qwen3-VL-32B as a larger-model reference. All models are evaluated using the same radar observations, task instructions, and decoding protocol.

Evaluation metrics. Radar-domain correctness is evaluated along five dimensions: motion direction, echo intensity, structure, evolution trend, and temporal consistency. Motion evaluates whether the dominant displacement direction of the radar echoes is correctly identified; intensity evaluates whether the strength of the observed echoes is correctly described; structure evaluates whether the major echo morphology and spatial configuration are correctly characterized; evolution trend evaluates whether the model correctly identifies the types and directions of radar-echo changes, such as strengthening, weakening, expansion, contraction, merging, splitting, and dissipation; and temporal consistency evaluates whether these changes are assigned to the correct stages of the radar sequence and described in the correct chronological order.

For each event i and evaluation dimension d, multiple domain experts independently assign a correctness score

$$r_{i,d}^{(m)} \in \{0, 0.5, 1\}, \tag{8}$$

where m indexes the expert. A score of 1 denotes a correct and sufficiently complete description, 0.5 denotes a partially correct but incomplete description, and 0 denotes an incorrect, unsupported, or completely missing description for that dimension.

The event-level score is obtained by averaging the judgments of the M domain experts:

$$s_{i,d} = \frac{1}{M} \sum_{m=1}^{M} r_{i,d}^{(m)}. \tag{9}$$

The score for each dimension is then averaged over all evaluated events and reported as a percentage:

$$S_d = \frac{100}{N} \sum_{i=1}^{N} s_{i,d}, \tag{10}$$

where $d \in \{\text{motion}, \text{intensity}, \text{structure}, \text{evolution}, \text{temporal}\}$.

We report the arithmetic mean of the five domain-specific scores as the overall *Radar Score*,

$$\text{Radar Score} = \frac{S_{\text{motion}} + S_{\text{intensity}} + S_{\text{structure}} + S_{\text{evolution}} + S_{\text{temporal}}}{5}. \tag{11}$$

Radar Score is used as the aggregate metric, while all five component scores are also reported to expose differences in specific radar-understanding capabilities. We do not use lexical-overlap metrics such as ROUGE-L, CIDEr, or BERTScore as primary evaluation measures, since lexical similarity does not directly measure the observational correctness of radar-event interpretation.

As a supplementary automated evaluation, we additionally use GPT-5.6 Sol as an independent evaluator under the same five radar-understanding dimensions and the same three-level scoring rubric $\{0, 0.5, 1\}$. GPT-5.6 Sol evaluates each model output with respect to motion, intensity, structure, evolution trend, and temporal consistency. The resulting scores are averaged across events and dimensions to obtain the *GPT Score*. We report only the aggregate GPT Score in the main experimental tables, while the corresponding dimension-level GPT evaluation results are provided in the appendix. GPT Score is used as an auxiliary evaluation and does not replace the domain-expert-based Radar Score.

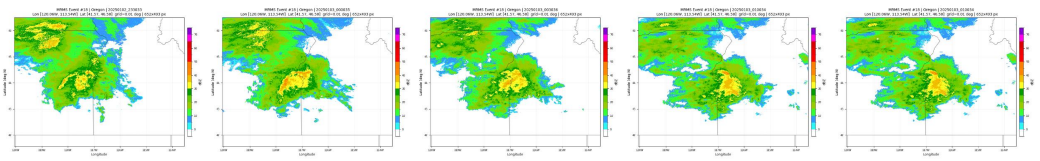
Controlled fine-tuning. All supervision-quality comparisons use the same Qwen3-VL-8B backbone. The accepted training-set size, LoRA configuration, optimizer, training steps, image/frame budget, decoding settings, and model-selection procedure are held fixed across variants. Therefore, the supervision-ablation experiments differ only in the verification stage applied to the generated supervision.

Unless otherwise specified, LoRA adapters are applied to both the attention projection layers and the MLP projection layers of Qwen3-VL-8B. The influence of the LoRA target modules is studied separately in Section 3.5.

3.2 Main Results

We first evaluate whether reliable expert-verified supervision can specialize a general-purpose VLM for radar-event understanding. Table 1 compares RadarVL-8B with its untuned Qwen3-VL-8B backbone, three general-purpose VLMs of comparable parameter scale, and the larger Qwen3-VL-32B model.

Prompt : *Act as a meteorological expert with strong experience in severe convective radar analysis and produce event-level evolution analysis from time-sorted radar composite reflectivity frames.*



Result : *This event depicts the development and evolution of a continuously eastward-moving banded echo system over central Oregon. In the initial stage, the echo system spanned from central to northern Oregon. Its main reflectivity remained between 20–30 dBZ, with localized intense cores reaching 30–40 dBZ. Meanwhile, an independent strong echo center existed to the north of the system, with reflectivity exceeding 40 dBZ near approximately 119.0 °W.Subsequently, the system propagated eastward as a whole and intensified gradually. By 00:30:36 UTC on 3 January 2025, the reflectivity of the primary intense core increased to 40–50 dBZ. The echo assumed an irregular clustered shape without a distinct linear orientation, featuring continuous and stable organization and attaining its maximum areal coverage, indicating the system's peak development at this stage .During the subsequent period from 01:00:34 UTC to 02:00:35 UTC on 3 January 2025, the intensity of the primary intense core decreased to 30–40 dBZ while its position continued shifting eastward. The echo pattern gradually dispersed with a slight reduction in spatial extent, and the overall system entered a weakening phase. The independent strong echo system on the northern side weakened and dissipated during this interval; no prominent merging or splitting processes were observed, and the dominant echo system maintained continuous evolutionary characteristics.*

Figure 2: Results Presentation

The untuned Qwen3-VL-8B obtains a Radar Score of 64.8, indicating that a general-purpose VLM has limited capability for directly interpreting continuous radar evolution. Other general-purpose models of similar scale provide moderate improvements, reaching Radar Scores between 65.9 and 68.0. In contrast, RadarVL-8B achieves 77.1, substantially improving all five radar-specific dimensions over the similarly sized baselines.

Table 1: Main results on the held-out MRMS event test set. Radar-specific scores are percentages and higher is better. Radar Score is based on domain-expert evaluation, while GPT Score provides a supplementary automated evaluation. Bold denotes the best overall result.

Model	Scale	Motion	Intensity	Structure	Evolution	Temporal	Radar Score	GPT Score
Qwen3-VL-8B	8B	63.2	66.1	65.5	62.8	66.4	64.8	61.0
InternVL3.5-8B	8.5B	65.0	66.9	66.4	64.2	67.0	65.9	64.3
MiniCPM-V 4.5	8B	66.4	67.8	67.1	65.7	67.5	66.9	66.1
GLM-4.1V-9B-Thinking	9B	67.5	68.9	68.2	66.6	69.0	68.0	70.2
Qwen3-VL-32B	32B	76.3	78.3	75.6	**76.8**	**76.9**	76.8	79.4
RadarVL-8B	8B	**77.4**	**78.6**	**76.2**	76.4	76.7	**77.1**	**80.6**

3.3 Effect of Supervision Verification

We next isolate the contribution of the supervision-construction procedure. Starting from the same Qwen3-VL-8B backbone, we progressively improve the training supervision while keeping the downstream fine-tuning configuration fixed. Specifically, we compare: (1) the untuned base model, (2) SFT directly using raw candidate annotations generated by Qwen3-VL-32B, (3) SFT using supervision after independent auditing and cross-verification, and (4) SFT using the final supervision after expert verification.

Direct SFT on raw candidate annotations increases Radar Score from 64.8 to 69.8, showing that the candidate annotations generated by Qwen3-VL-32B already contain useful radar-domain information.

Table 2: Effect of supervision verification on downstream radar understanding. All SFT variants use the same Qwen3-VL-8B backbone, matched training-set size, and identical training configuration.

Training setting	Motion	Intensity	Structure	Evolution	Temporal	Radar Score	GPT Score
Qwen3-VL-8B (no SFT)	63.2	66.1	65.5	62.8	66.4	64.8	61.0
Raw candidate annotation SFT	68.6	71.0	70.2	68.4	70.8	69.8	67.4
Audit + cross-verified supervision SFT	75.8	73.5	72.1	74.0	73.9	73.9	75.6
Audit + cross-verification + expert verification	**77.4**	**78.6**	**76.2**	**76.4**	**76.7**	**77.1**	**80.6**

However, substantially larger gains are obtained after reliability control. Independent auditing and cross-verification increase Radar Score to 73.9, with improvements across all five radar dimensions.

Final expert verification further increases the score from 73.9 to 77.1. The progressive improvement from raw supervision to automatically verified supervision and finally to expert-verified supervision indicates that downstream specialization benefits not only from exposure to radar-domain annotations, but also from improving the reliability of those annotations before they are used for SFT.

3.4 Analysis of Independent Auditing

A central design choice in our supervision-construction pipeline is that the independent auditor analyzes the original radar observations without access to the candidate annotation. We investigate whether this information independence is necessary by comparing candidate-aware and candidate-blind auditing under the same cross-verification procedure.

We randomly sample 500 events with raw candidate annotations generated by Qwen3-VL-32B for this analysis. The same sampled events are used for both auditing protocols. In the *candidate-aware* setting, the auditor receives both the original radar sequence and the generated candidate annotation. In the *candidate-blind* setting, the auditor receives only the radar sequence and independently constructs an observation-based interpretation before cross-verification.

For evaluation, candidate annotations are decomposed into atomic scientific claims. After cross-verification, domain experts independently inspect the remaining claims against the original radar observations and determine whether each claim is correct. Expert judgments are used only to evaluate the auditing results in this experiment; no expert correction is applied and no downstream fine-tuning is performed.

The raw candidate annotations in the sampled subset have an atomic-claim error rate of 17.9%. Table 3 reports the absolute reduction in error rate after auditing and cross-verification.

Table 3: Atomic-claim error rates before and after different auditing protocols on the same 500 sampled candidate events.

Protocol	Atomic-claim error rate (%)↓
Raw candidate	17.9
Candidate-aware audit	10.6
Candidate-blind audit	**7.8**

Candidate-aware auditing reduces the atomic-claim error rate by 7.3 percentage points, whereas candidate-blind auditing achieves a larger reduction of 10.1 percentage points under the same cross-verification procedure.

This difference supports the use of information-independent auditing. When the candidate annotation is visible during auditing, the auditor may be influenced by the framing, interpretation, or intermediate conclusions already present in the generated result, making existing errors more likely to be inherited. Candidate-blind auditing instead requires the auditor to first form an independent interpretation directly from the radar observations and only then compare it with the candidate during cross-verification, thereby providing a more independent source of evidence for identifying erroneous claims.

3.5 Ablation of LoRA Target Modules

We further investigate which parts of the language backbone should be adapted during radar-domain specialization. Our default configuration inserts LoRA adapters into both the attention projection layers and the MLP projection layers of Qwen3-VL-8B. We compare this configuration with variants in which LoRA is applied only to the attention projections or only to the MLP projections. All variants use exactly the same expert-verified supervision and training configuration.

Table 4: Ablation of LoRA target modules. All variants are trained using the same expert-verified supervision.

LoRA target modules	Motion	Intensity	Structure	Evolution	Temporal	Radar Score	GPT Score
Attention only	75.1	76.3	74.4	73.2	75.5	74.9	76.8
MLP only	74.3	75.8	74.8	73.7	75.4	74.8	75.1
Attention + MLP	**77.4**	**78.6**	**76.2**	**76.4**	**76.7**	**77.1**	**80.6**

Applying LoRA only to the attention projections or only to the MLP projections already provides substantial radar-domain adaptation, reaching Radar Scores of 74.9 and 74.8, respectively. However, adapting both components produces the strongest overall result, increasing Radar Score to 77.1 and improving all five radar-specific dimensions.

These results suggest that radar-event specialization benefits from jointly adapting the attention-based interaction mechanism and the feed-forward representations of the language backbone, rather than restricting low-rank adaptation to either component alone.

3.6 Zero-Shot Generalization to Unseen OPERA Radar Data

Finally, we evaluate whether specialization on MRMS observations transfers to radar data from a different product and geographic region. RadarVL-8B is trained exclusively using MRMS-derived Radar-9K supervision and is directly evaluated on 400 unseen European OPERA events without any OPERA fine-tuning or adaptation.

Table 5: Zero-shot evaluation on 400 unseen OPERA radar events. No OPERA observations are used during supervision construction or fine-tuning.

Model	Motion	Intensity	Structure	Evolution	Temporal	Radar Score	GPT Score
Qwen3-VL-8B	60.9	63.1	62.0	60.2	62.3	61.7	58.9
Qwen3-VL-32B	69.7	72.4	71.1	70.3	70.5	70.8	72.7
RadarVL-8B	**71.8**	**72.9**	**71.5**	**70.7**	**70.6**	**71.5**	**74.6**

The untuned Qwen3-VL-8B obtains a Radar Score of 61.7 on OPERA, indicating a substantial domain gap for direct general-purpose VLM inference. Despite never observing OPERA data during training, RadarVL-8B improves the score to 71.5, a gain of 9.8 points over its untuned backbone. It also remains competitive with, and slightly exceeds, the larger Qwen3-VL-32B model.

The improvement is observed across all five radar-understanding dimensions, suggesting that the learned radar-domain capability is not restricted to the specific MRMS observations used for specialization. This experiment therefore provides evidence of transfer across radar products and geographic regions within the same scientific domain, while not implying unrestricted cross-domain generalization.

4 Discussion and Limitations

Verified supervision matters. The experiments show that strong VLMs can provide useful radar-domain candidate annotations, but their raw outputs should not be directly regarded as reliable training labels. Direct SFT on raw candidate annotations already improves the base model, indicating that the generated candidates contain useful domain information. However, substantially stronger performance is obtained after independent auditing, cross-verification, and expert verification. This suggests that the benefit of model-generated supervision depends not only on the knowledge contained

in the generated annotations, but also on whether their claims are sufficiently verified against the original observations before being used for specialization.

Why candidate-blind auditing matters. The auditing experiment further shows that information independence is important during supervision verification. Under the same cross-verification procedure, candidate-blind auditing produces a larger reduction in atomic-claim error rate than candidate-aware auditing. When the candidate annotation is visible, the auditor may be influenced by its existing framing, interpretations, or intermediate conclusions. Requiring the auditor to first interpret the radar observations independently therefore provides a more independent source of evidence for subsequent verification. Nevertheless, different VLMs may still share training data, architectures, or systematic biases, so such cross-model auditing should be understood as information-level independence rather than statistical independence in the sense of independent human annotators.

Scope and limitations. Our empirical evidence is restricted to weather radar, with supervision constructed from 2025 MRMS composite-reflectivity observations and external evaluation conducted on OPERA composite-reflectivity observations. The results therefore support reliability improvements in radar supervision and within-domain transfer across radar products and geographic regions, rather than unrestricted cross-domain generalization. Expert judgment remains necessary both for final supervision verification and for domain-specific evaluation, which introduces unavoidable human assessment into the pipeline. In addition, processing long multimodal sequences and performing multiple VLM inference passes increase computational cost. Ambiguity may also originate upstream from event tracking, particularly during complex splitting, merging, and rapid structural reorganization. Whether the same verification principle extends to other scientific instruments and professional domains remains an open empirical question.

5 Related Work

General VLMs and visual instruction tuning have established strong foundations for multimodal adaptation (Alayrac et al., 2022; Li et al., 2023b; Dai et al., 2023; Liu et al., 2023; Lin et al., 2024). Domain specialization has been explored in biomedicine (Li et al., 2023a; Moor et al., 2023; Nath et al., 2025), remote sensing (Kuckreja et al., 2024; Muhtar et al., 2024; Zhang et al., 2024), agriculture (Awais et al., 2025), and ocean science (Bi et al., 2024). Meteorological and radar applications have similarly advanced from storm identification and tracking (Dixon & Wiener, 1993; Heikenfeld et al., 2019; Sokolowsky et al., 2024; Feng et al., 2023) and precipitation nowcasting (Shi et al., 2015; Ravuri et al., 2021; Zhang et al., 2023; Gao et al., 2023; Gong et al., 2024) to multimodal weather understanding and report generation, including WeatherQA (Ma et al., 2024), RadarQA (He et al., 2025), and recent weather and climate reasoning studies (Li et al., 2025; Higgins et al., 2026; Zheng et al., 2026).Synthetic instructions and VLM-generated rationales can reduce supervision cost (Wang et al., 2023; Ho et al., 2023; Hsieh et al., 2023; Cai et al., 2025), but model-generated data may preserve or amplify errors without verification (Feng et al., 2025). Related work therefore studies multimodal hallucination and self-correction (Li et al., 2023c; Guan et al., 2024; Jing et al., 2024; Liu et al., 2025). Our work focuses on constructing reliable language supervision for complete radar events by verifying VLM-generated professional claims against source observations, with candidate-blind auditing and expert verification.

6 Conclusion

We study how candidate annotations generated by a strong VLM can be converted into reliable supervision for continuous weather-radar understanding. Rather than directly treating generated outputs as training labels, we verify them through explicit radar observation boundaries, independent candidate-blind auditing, observation-grounded cross-verification, and expert review. The resulting Radar-9K supervision improves the Radar Score of Qwen3-VL-8B from 64.8 to 77.1, while candidate-blind auditing removes more atomic-claim errors than candidate-aware auditing. The resulting RadarVL-8B also transfers without additional fine-tuning to unseen OPERA radar observations, achieving a Radar Score of 71.5. These results show that strong VLMs can provide useful candidate supervision for radar specialization, while reliability verification is important before model-generated professional annotations are used as training labels.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and improving the readability of the manuscript, as well as to support literature retrieval and discovery of relevant prior work. All scientific content, methodological claims, experimental results, and conclusions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the main methodological and experimental details required for reproducibility in the paper and appendix. Section 2 describes the supervision-construction pipeline and dataset construction, while Section 3 presents the experimental and evaluation protocols. Additional details on task prompts and radar observation boundaries, radar-event segmentation, implementation settings, qualitative evaluation examples, and GPT-5.6 Sol evaluation are provided in Appendices, respectively. Code and processed metadata will be released subject to the redistribution policies and licenses of the underlying MRMS and OPERA radar products.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. AgroGPT: Efficient agricultural vision-language model with expert tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Awais_AgroGPT_Efficient_Agricultural_Vision-Language_Model_with_Expert_Tuning_WACV_2025_paper.html.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. OceanGPT: A large language model for ocean science tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3357–3372. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.184. URL https://aclanthology.org/2024.acl-long.184/.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. LLaVA-KD: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 239–249, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cai_LLaVA-KD_A_Framework_of_Distilling_Multimodal_Large_Language_Models_ICCV_2025_paper.html.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract.html.

Michael Dixon and Gerry Wiener. TITAN: Thunderstorm identification, tracking, analysis, and nowcasting—a radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, 10(6):785–797, 1993. doi: 10.1175/1520-0426(1993)010⟨0785:TTITAA⟩2.0.CO;2. URL https://impacts.ucar.edu/en/publications/titan-thunderstorm-identification-tracking-analysis-and-nowcastin/.

EUMETNET OPERA. OPERA CIRRUS Maximum Reflectivity Composite: User Product Sheet. Maximum Reflectivity Composite, 1 km / 5 min, Edition 2.0, 2024. URL https://www.eumetnet.eu/wp-content/uploads/2024/06/OPERA_Max-Reflectivity_Product-Sheet_Ed-2.0.pdf.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, François Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df2d62b96a4003203450cf89cd338bb7-Abstract-Conference.html.

Zhe Feng, L. Ruby Leung, Joseph Hardin, and Christopher R. Terai. PyFLEXTRKR: A flexible feature tracking python software for convective cloud analysis. *Geoscientific Model Development*, 16:2753–2776, 2023. doi: 10.5194/gmd-16-2753-2023. URL https://doi.org/10.5194/gmd-16-2753-2023.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle C. Maddix, Yan Zhu, Mu Li, and Yuyang Wang. PreDiff: Precipitation nowcasting with latent diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f82ba6a6b981fbbecf5f2ee5de7db39c-Abstract.html.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, et al. GLM-4.1V-Thinking and GLM-4.5V: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006.

Junchao Gong et al. CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024. URL https://arxiv.org/abs/2402.04290.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/beb667eee2fecb594dc3154950ef7aa4-Abstract-Conference.html.

Max Heikenfeld, Peter J. Marinescu, Matthew Christensen, Duncan Watson-Parris, Fabian Senf, Susan C. van den Heever, and Philip Stier. tobac 1.2: Towards a flexible framework for tracking and analysis of clouds in diverse datasets. *Geoscientific Model Development*, 12(11):4551–4570, 2019. doi: 10.5194/gmd-12-4551-2019. URL https://gmd.copernicus.org/articles/12/4551/2019/.

Timothy B. Higgins, Antonios Mamalakis, and Chirag Agarwal. SynopticBench: Evaluating vision-language models on generating weather forecast discussions of the future. *arXiv preprint arXiv:2604.16451*, 2026. URL https://arxiv.org/abs/2604.16451.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. URL https://arxiv.org/abs/2305.02301.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kuckreja_GeoChat_Grounded_Large_Vision-Language_Model_for_Remote_Sensing_CVPR_2024_paper.html.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 36, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Hobo Li et al. CLLMate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17536–17562, 2025. doi: 10.18653/v1/2025.emnlp-main.886. URL https://aclanthology.org/2025.emnlp-main.886/.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023c. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_VILA_On_Pre-training_for_Visual_Language_Models_CVPR_2024_paper.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Zixuan Liu, Guangkai Jiang, and Siavash Khajavi. LLaVA-SCo: Teach vision language models to self-correct. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3445–3454, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/WiCV/html/Liu_LLaVA-SCo_Teach_Vision_Language_Models_to_Self-Correct_CVPRW_2025_paper.html.

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. URL https://arxiv.org/abs/2406.11217.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: A multimodal medical few-shot learner. In *Proceedings of Machine Learning for Health*, volume 225 of *Proceedings of Machine Learning Research*, 2023. URL https://arxiv.org/abs/2307.15189.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xuelong Zhang, and Pengfeng Xiao. LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model. In *Computer Vision – ECCV 2024*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09511.pdf.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. VILA-M3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Nath_VILA-M3_Enhancing_Vision-Language_Models_with_Medical_Expert_Knowledge_CVPR_2025_paper.pdf.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672–677, 2021. doi: 10.1038/s41586-021-03854-z. URL https://www.nature.com/articles/s41586-021-03854-z.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/5955-convolutional-lstm-network-a-machine-learningapproach-for-precipitation-nowca pdf.

G. A. Sokolowsky et al. tobac v1.5: Introducing fast 3-d tracking, splits and mergers, and other enhancements for identifying and analysing meteorological phenomena. *Geoscientific Model Development*, 17:5309–5330, 2024. doi: 10.5194/gmd-17-5309-2024. URL https://gmd.copernicus.org/articles/17/5309/2024/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 13484–13508, 2023. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Han Xu, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. URL https://arxiv.org/abs/2509.18154.

Wei Zhang et al. EarthGPT: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024. doi: 10.1109/TGRS.2024.3409624. URL https://arxiv.org/abs/2401.16822.

Yuchen Zhang, Mingsheng Long, Kai Chen, Lanxiang Xing, Rong Jin, Michael I. Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619: 526–532, 2023. doi: 10.1038/s41586-023-06184-4. URL https://www.nature.com/articles/s41586-023-06184-4.

Zinan Zheng, Yang Liu, Nuo Chen, Juepeng Zheng, Hong Cheng, and Jia Li. WeatherSyn: An instruction tuning MLLM for weather forecasting report generation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=WlScVvMwpB.

A TASK PROMPTS, SFT INSTRUCTION, AND RADAR OBSERVATION BOUNDARIES

This appendix provides the prompts used for candidate generation and candidate-blind independent auditing, the instruction format used for supervised fine-tuning (SFT), and the radar observation boundaries used throughout the supervision-construction process. All prompts operate on temporally ordered composite-reflectivity observations and follow the same observation-grounded interpretation principles described in the main paper.

A.1 PROMPT A: CANDIDATE-BLIND INDEPENDENT AUDIT

Prompt A is used for candidate-blind independent auditing. The auditor receives only the temporally ordered radar observations and independently records observation-supported radar facts without access to the candidate annotation. The resulting structured audit provides an independent observational reference for subsequent cross-verification.

```
### Role

You are an independent severe-convective radar fact-audit expert. You
    cannot see any candidate description and must directly inspect all
    two-dimensional CREF images arranged in chronological order. Output
    an independent visual record for final review. Do not write a
    complete event narrative.

### Audit Rules

* Distinguish coexisting echo systems within the bounding box. Record
    objects as the same system only when spatial proximity, morphological
     continuity, and temporal continuity are all simultaneously supported.
     Otherwise, record the relationship as uncertain or unrelated. Set
    multiple_systems=true when multiple systems coexist.
* Cite a Frame for every observation and use verifiable longitude/
    latitude coordinates or clear administrative locations. Check the
    color scale frame by frame. When intensity boundaries cannot be
    reliably distinguished, use a broader dBZ range or describe only the
    relative change.
* Maintain a falsification-oriented approach when verifying movement,
    connection, merging, splitting, transitions in the dominant center,
    and peak maintenance. Record the relationship as uncertain when the
    evidence is insufficient.
* When verifying a diagonal principal axis, independently read the
    longitude and latitude of the western and eastern endpoints. An
    increase in longitude accompanied by a decrease in latitude indicates
     a northwest-southeast orientation. An increase in both longitude and
     latitude indicates a southwest-northeast orientation. If the
    coordinates or principal axis are unclear, record the orientation as
    uncertain.
* Record only facts directly visible in two-dimensional CREF images. Do
    not record color-based descriptions, visually estimated distances or
    areas, vertical structure, velocity products, environmental
    mechanisms, or surface hazards.
* Inspect all frames. When consecutive states are similar, merge them
    using a Frame range to avoid redundant frame-by-frame descriptions or
     incomplete coverage.

### Output

Output only one parseable JSON object. Do not output Markdown,
    explanations, or a written event narrative.
```

A.2 PROMPT B: CANDIDATE ANNOTATION GENERATION

Prompt B is used to generate the event-level candidate annotation. The model directly analyzes the complete temporally ordered radar sequence and produces a structured, observation-grounded representation of the event before the generated result enters the subsequent verification process.

```
### Role and Input

You are a meteorological expert with experience in severe-convective
    radar analysis. Use process-analysis language consistent with
    meteorological research papers to conduct a professional, coherent,
    and verifiable event-level analysis of composite reflectivity images
    arranged in chronological order. The input images belong to the same
    event. Each image has a unique Frame number and an accurate timestamp.


Directly inspect all images and output an event-level, evidence-
    constrained, manually verifiable intermediate JSON.

The intermediate JSON uses only Frame numbers. After programmatic
    validation, all Frame fields and textual references will be
    deterministically converted into the exact timestamps encoded in the
    PNG filenames. The final training JSON will not retain Frame numbers.

### Evidence-First Workflow

Complete the following internally without outputting the reasoning
    process:

1. Compare the intensity, strong cores, spatial extent, morphology,
    organization, and position across consecutive frames.
2. Distinguish coexisting echo systems within the same bounding box and
    determine the dominant object during each period.
3. Before writing, determine the overall event structure: whether it
    represents the continuous evolution of a single echo system, multiple
     relatively independent processes occurring in parallel, or a
    transition in the dominant object.
4. Divide the event into stages only when substantive changes occur, and
    determine the peak and terminal trend.
5. After generation, verify every statement against the images frame by
    frame. Remove unsupported claims, incorrect Frame references,
    spatially inconsistent descriptions, and conclusions beyond what can
    be established from two-dimensional composite reflectivity images.

### Target Identity

* A continuous movement may be described as involving the same object
    only when adjacent frames simultaneously support spatial proximity,
    morphological continuity, and temporal continuity.
* When continuous echo evidence is missing or targets are spatially
    distant, do not claim connection, merging, splitting, replacement, or
     the same lifecycle. Describe the objects separately or omit the
    relationship.
* When multiple systems coexist, do not force them into a single system
    merely to construct a complete narrative.

### Spatial and Orientation Grounding

* Absolute locations must use verifiable administrative place names or
    longitude/latitude coordinates from the basemap. When a place name is
     unclear, use coordinates only and do not guess.

* Do not use vague expressions such as "the western part of the image", "
    the middle of the frame", or "the northern part of the study area".
    Relative positions within an identified object are allowed, such as "
```

```
    the northeastern side of the main echo", but such expressions cannot
    replace an absolute geographic anchor.

* Whenever describing a northwest-southeast or southwest-northeast
    principal axis, the same sentence must provide both endpoints in the
    following parseable format:

  "the western end of the principal axis is approximately at longitude
    degrees E, latitude degrees N, and the eastern end is approximately
    at longitude degrees E, latitude degrees N, forming a [direction]
    orientation"

* The longitude of the eastern endpoint must be greater than that of the
    western endpoint.

* A decrease in latitude toward the eastern endpoint corresponds to a
    northwest-southeast orientation; an increase in latitude corresponds
    to a southwest-northeast orientation.

* If the two endpoints or the principal axis cannot be determined clearly,
     do not describe a diagonal orientation. The stage_name must not
    contain an orientation that is not supported by coordinates.

### Stage and Frame Rules

* Output at least one stage. There is no upper limit on the number of
    stages.
* Consecutive similar frames should be grouped into the same stage. Do
    not mechanically create one stage per frame, and do not compress a
    complex long-duration event into a fixed number of stages.
* A new stage may be established when one or more of the following
    evidence-supported substantive changes occur:

  * clear strengthening, weakening, expansion, or contraction of the echo
    ;
  * formation, disappearance, change in number, or change in dominance of
     high-reflectivity cores;
  * merging, splitting, replacement, or emergence of a new echo that
    gradually becomes dominant;
  * a clear change in the planar morphology, organizational continuity,
    or movement direction of the main echo.
* Stages must continuously cover the sequence from Frame 1 through the
    final frame, without overlap or gaps. The input may cover only part
    of the lifecycle; do not invent unobserved stages.
* evidence_frames should contain only representative frames that directly
     support the conclusion.
* Every observable_evidence item must explicitly cite a Frame contained
    in that stage's evidence_frames.
* Frame numbers are used only for key evidence and transitions, not for
    listing every frame.
* Do not use vague temporal expressions such as "the early part of the
    sequence", "the later input", or "the image sequence".

### Observable Evidence Boundary

* Intensity must be expressed using dBZ values, thresholds, or ranges
    directly supported by the image color scale. Under no circumstances
    may color names, color-block names, or colors used as substitutes for
     reflectivity intensity be output.
* When color-scale boundaries cannot be reliably distinguished, do not
    describe colors. Use a broader dBZ range directly supported by the
    scale, or describe only relative strengthening or weakening without
    guessing precise numerical values.
* Without external computational results, do not estimate distances,
    areas, or length/width dimensions by visual inspection.
```

```
* Do not infer echo-top height, vertical development, BWER, dome-like
    structures, overhangs, sloping ascent, low-level extension, or ground
     contact.
* Do not infer updrafts or downdrafts, velocity couplets, convergence,
    divergence, wind shear, mesocyclones, tornadoes, thermodynamic or
    dynamical environmental conditions, moisture, triggering mechanisms,
    or confirmed surface hazards such as hail, precipitation, heavy
    rainfall, or strong winds.

### Next-Stage Outlook

stages, evolution_analysis, and peak_stage describe only observed facts.
    next_stage_outlook is the only field that permits extrapolation.

* trend must be one of: "strengthening", "weakening", or "maintaining".
* expected_evolution must begin with "Expected..." and describe only one
    or two changes most strongly supported by the terminal evidence. Do
    not use expressions such as "may", "might", "possibly", "perhaps", "
    likely", "suspected", or "could".
* basis_frames must include the final Frame. When the input contains more
     than one frame, it must contain at least two different Frames,
    preferably from the final stage.
* confidence must be one of: "high", "medium", or "low".
* uncertainty must cite specific basis_frames to explain the boundary of
    the observable evidence.
* Do not invent future Frames, timestamps, precise locations, precise dBZ
     values, or specific hazards.

### Output

Output only one parseable JSON object.

Do not output Markdown, explanations, introductions, code fences,
    reasoning processes, or self-generated timestamp strings.
```

A.3 SFT Instruction Format

After expert verification, the final annotation Y_i^* is reformulated as multimodal instruction data for supervised fine-tuning of Qwen3-VL-8B. Each training example consists of an ordered radar-event sequence, a radar-analysis instruction, and the corresponding expert-verified annotation.

Conceptually, an SFT example is represented as

$$\mathcal{D}_{\mathrm{SFT}} = \left\{ \left(\mathcal{R}_i, q_i^{\mathrm{SFT}}, Y_i^* \right) \right\}_{i=1}^{N}, \tag{12}$$

where $\mathcal{R}_i$ is the temporally ordered radar sequence, q_i^{SFT} is the task instruction, and Y_i^* is the final expert-verified radar annotation.

The instruction used in the SFT samples follows the template below.

```
Act as a meteorological expert with strong experience in severe
    convective radar analysis and produce event-level evolution analysis
    from time-sorted radar composite reflectivity frames.
```

The assistant response contains only the corresponding expert-verified annotation Y_i^* rather than intermediate auditing or verification information.

A.4 Radar Observation Boundaries

All stages of supervision construction follow the radar observation boundaries introduced in Section 2.2:

$$\mathcal{K} = \{\mathcal{K}_{\mathrm{obs}}, \mathcal{K}_{\mathrm{rel}}, \mathcal{K}_{\mathrm{forbid}}\}. \tag{13}$$

These boundaries are summarized in Table 6.

Table 6: Radar observation boundaries used during candidate generation, independent auditing, cross-verification, and expert verification.

Boundary	Interpretation
$\mathcal{K}_{\mathrm{obs}}$	$\mathcal{K}_{\mathrm{obs}}$ covers directly observable properties including reflectivity intensity, echo location and extent, morphology, and organization.
$\mathcal{K}_{\mathrm{rel}}$	$\mathcal{K}_{\mathrm{rel}}$ covers relations supported by the temporal sequence, including translation, strengthening, weakening, expansion, contraction, merging, and dissipation.
$\mathcal{K}_{\mathrm{forbid}}$	$\mathcal{K}_{\mathrm{forbid}}$ excludes conclusions that cannot be reliably established from composite reflectivity alone, such as radial velocity, vertical structure, thermodynamic causes, surface damage, or direct confirmation of hail and tornado occurrence.

These constraints apply to the scientific content of the annotation rather than merely to its linguistic form. A fluent statement is therefore not considered valid if the corresponding conclusion cannot be supported by the original radar observations.

B Radar Event Segmentation: Parameters and Design Details

The main paper describes radar-event construction at a conceptual level. This appendix provides additional implementation details for converting continuous radar observations from the MRMS and OPERA products into temporally coherent convective events.

The event-construction procedure consists of four main stages: (1) strong-echo cell detection, (2) within-frame spatial clustering, (3) online temporal tracking, and (4) lifecycle expansion and spatial context completion. The central design principle is to distinguish the identification of a strong convective event from the recovery of its broader temporal and spatial lifecycle. Relatively strict criteria are therefore used to establish the event identity, while a lower reflectivity threshold is introduced only after a valid event has been established.

B.1 Strong-Echo Cell Detection

For each radar frame, candidate convective cells are first extracted using a reflectivity threshold of 35 dBZ. Binary closing with a 3×3 structuring element is applied before connected-component labeling to reduce small discontinuities within otherwise contiguous echo regions. Connected components containing fewer than 80 pixels are discarded.

The 35 dBZ threshold is intended to identify the primary convective structure rather than all precipitation echoes. A substantially lower threshold would introduce a larger number of weak and stratiform echoes into the event-definition stage, increasing the number of irrelevant candidate components. In contrast, an excessively high threshold could retain only the strongest portions of a convective system and fragment its spatial representation. The selected threshold therefore provides an intermediate criterion for extracting the strong echo structures used for subsequent tracking.

The minimum cell area of 80 pixels provides an additional noise-suppression mechanism. Very small connected components are more likely to correspond to isolated weak echoes or fragmented structures and are therefore removed before spatial clustering. This reduces the number of spurious candidates while retaining sufficiently large convective cells.

B.2 Within-Frame Spatial Clustering

Multiple detected cells in the same radar frame may represent different convective cores belonging to the same organized system. We therefore group spatially nearby cells using DBSCAN with an 80 km neighborhood radius. The distance between cell centroids is computed using the haversine metric, so that the clustering criterion is expressed directly in physical distance rather than in image pixels.

DBSCAN is suitable for this stage because the number of convective systems and their spatial organization vary between radar observations. It does not require a predefined number of clusters and can naturally group multiple nearby convective cores into one spatially organized system.

The 80 km radius is intended to provide a sufficiently broad association scale for nearby convective cells while avoiding the indiscriminate merging of geographically separated systems. This stage therefore transforms individual thresholded cells into spatially organized clusters that provide a more stable representation for temporal tracking.

B.3 Online Temporal Tracking

After spatial clustering, clusters are associated across consecutive radar observations using an online tracking procedure. Instead of constructing a global temporal graph over the entire radar sequence, the implementation maintains active tracks and performs frame-to-frame matching. This streaming design reduces memory consumption while preserving local temporal continuity.

For two candidate clusters in consecutive observations, the matching score is computed as

$$S = 0.35S_{\mathrm{IoU}} + 0.25S_{\mathrm{speed}} + 0.10S_{\mathrm{direction}} + 0.15S_{\mathrm{area}} + 0.15S_{\mathrm{intensity}}. \tag{14}$$

The individual terms characterize complementary properties of temporal continuity. The intersection-over-union term receives the largest weight (0.35), since spatial overlap provides direct evidence that two clusters represent the same evolving echo structure. The speed-related term receives a weight of 0.25 and evaluates whether the observed displacement is consistent with plausible system motion. The direction term receives a smaller weight of 0.10, because convective systems may deform, merge, split, or change propagation direction during their evolution.

Area and reflectivity continuity are assigned equal weights of 0.15. Area similarity provides a structural constraint, while reflectivity similarity provides an additional physical consistency signal. Combining these terms avoids relying on a single geometric criterion when matching complex or evolving convective structures.

The default matching threshold is 0.35. Candidate pairs whose centroid distance exceeds 300 km are rejected before the composite matching score is evaluated. This provides a two-stage association mechanism: an initial geographical gate removes clearly implausible matches, while the composite score evaluates the consistency of the remaining candidates.

The matching procedure is one-to-one within each temporal transition. Unmatched clusters can initialize new tracks, while active tracks that fail to obtain a valid continuation are terminated according to the configured gap constraint. The default maximum tracking gap is zero, meaning that the primary event trajectory is required to remain continuous between adjacent observed frames.

B.4 Anchor-Based Event Definition

Not every temporally tracked object is retained as a target convective event. A tracked system must contain at least one sufficiently large and intense anchor cluster satisfying

$$A \geq 2000 \text{ pixels}, \qquad \mathrm{maxDBZ} \geq 50 \text{ dBZ}. \tag{15}$$

The anchor criterion is intentionally stricter than the initial 35 dBZ cell-detection threshold. The purpose of the anchor stage is to distinguish the target convective systems from transient weak echoes and small isolated cells. Temporal continuity alone is therefore insufficient to define an event; the track must also contain a sufficiently strong and spatially substantial convective core.

An event must persist for at least four observed radar frames. With the nominal temporal sampling interval of 30 minutes used by the extraction pipeline, this corresponds to approximately 1.5 hours of observations. This duration constraint removes very short-lived detections that are less suitable for describing meaningful temporal evolution.

The separation between the initial detection threshold and the anchor criterion is important. The 35 dBZ threshold determines which structures enter the tracking process, whereas the 50 dBZ and 2000-pixel criteria determine whether a tracked structure is sufficiently strong and extensive to become an event.

B.5 Temporal Lifecycle Expansion

The strong-core event definition does not necessarily cover the complete observed lifecycle of a convective system. A system may initially appear as a weak echo, intensify into a strong convective core, and subsequently weaken below the threshold used during the primary event-definition stage.

To recover these developing and decaying stages, the established event is subsequently extended using a lower reflectivity threshold of 18 dBZ. This lower threshold is not used to independently create new events. Instead, it is applied only after a valid strong-core track has been identified.

Starting from the temporal boundaries of an established event, the algorithm searches backward and forward for weak-echo structures that remain spatially and temporally consistent with the existing event. The temporal search window is limited to six frames in each direction. With a nominal 30-minute temporal interval, this corresponds to a maximum extension of approximately three hours before the first strong-core observation and three hours after the final strong-core observation.

For temporal lifecycle matching, the implementation uses

$$S_{\text{temporal}} = 0.7S_{\text{IoU}} + 0.3S_{\text{distance}}, \tag{16}$$

where the distance term is normalized using a maximum centroid distance of 50 km. The larger IoU weight emphasizes structural continuity with the established event, while the distance term provides an additional constraint against associating unrelated weak echoes.

The temporal search is terminated after two consecutive frames without a valid match. This stopping rule prevents the lifecycle expansion from continuing indefinitely into weak background echoes that happen to occur near the previous event location.

B.6 Spatial Lifecycle Expansion

Temporal lifecycle expansion is complemented by a spatial expansion stage. Weak echoes surrounding the established strong-core event are identified using the same 18 dBZ contextual threshold. The search region is expanded by 200 pixels around the strong-event spatial extent before weak-echo components are evaluated.

The purpose of this stage is to recover peripheral and weak structures that are spatially connected to the established convective system but may not satisfy the strong-core criteria. Such structures can correspond to developing regions, decaying echoes, or weaker portions of an organized convective system.

Importantly, these weak-echo components do not redefine the event identity. Event identity is determined by the strong-core detection and temporal tracking stages. Lifecycle expansion only determines how much surrounding observational context is retained for the final radar-event sequence.

This separation is particularly important for downstream radar interpretation. Using the lower 18 dBZ threshold directly for event definition would substantially increase the number of weak or unrelated echoes that could be treated as independent events. Restricting the lower threshold to post-hoc contextual expansion instead preserves the selectivity of event identification while retaining useful lifecycle information.

B.7 Unified Spatial Extent and Rendering

For each established event, a unified spatial bounding box is computed across all selected observations. A margin of 120 pixels is added around the event extent to preserve surrounding radar context.

The margin is specified in pixels rather than geographic degrees because the pipeline is applied to radar products with their respective native grids, including MRMS and OPERA. This avoids assuming a common geographic resolution between the two products.

To prevent exceptionally large events from producing excessively large images, the radar data region is capped at 1400×1400 pixels. If the raw event extent exceeds this limit, the bounding box is centered on the strongest cluster and cropped to the maximum permitted dimensions. The strongest cluster is used as the spatial reference because it provides the most reliable indication of the primary convective structure.

The same bounding box is applied to every observation belonging to the same event. Consequently, the radar sequence remains spatially aligned across time, and changes between frames cannot be attributed to independent frame-level cropping.

The rendering stage preserves the native radar resolution. One radar data pixel is mapped to one rendered image pixel, without resizing, letterboxing, or padding of the radar data region. Annotation elements, including titles, coordinate labels, administrative boundaries, and colorbars, are placed outside the radar data region whenever possible. This design preserves the original spatial geometry and avoids introducing artificial spatial scaling during dataset construction.

B.8 Computational Constraints

Several implementation parameters are introduced for computational stability rather than for defining meteorological events. The DBSCAN operation uses single-process execution, while the underlying numerical libraries are restricted to a small number of threads. The lifecycle H5 cache is limited to four files, and the number of weak-echo cells processed per frame is capped at 2000.

These constraints are motivated by the substantially larger number of weak components that can be produced when the contextual threshold is reduced from the strong-core detection threshold to 18 dBZ. Without such limits, memory consumption and concurrent numerical operations can become unnecessarily large during large-scale processing.

The rendering stage similarly limits the number of concurrent workers to avoid excessive memory consumption when multiple high-resolution radar events are rendered simultaneously. These computational constraints do not change the event-definition criteria; they are engineering safeguards for stable large-scale dataset construction.

B.9 Summary of the Parameter Design

The parameterization follows a hierarchical event-construction strategy. First, a 35 dBZ threshold and an 80-pixel minimum area are used to identify sufficiently coherent echo cells. These cells are grouped using an 80 km spatial radius and associated temporally using a multi-component matching score. Tracks are then retained as target events only when they contain a strong anchor satisfying 2000 pixels and 50 dBZ and persist for at least four observations.

After event identity has been established, the lower 18 dBZ threshold is introduced only for lifecycle completion. Temporal expansion is limited to six frames in each direction and uses spatially constrained matching, while spatial expansion uses a 200-pixel contextual margin. Finally, a unified bounding box and native-resolution rendering strategy ensures that the complete event remains spatially aligned across time.

Overall, the design separates *event identity* from *event context*. Strong thresholds and temporal continuity provide a selective definition of the target convective event, whereas lower-threshold expansion recovers weaker structures that are useful for understanding its development and dissipation. The same extraction logic is applied to both MRMS and OPERA observations, while pixel-based parameters are interpreted relative to the native grid of each radar product.

C Limitations of Conventional NLP Metrics

Our task evaluates whether a generated radar interpretation is scientifically consistent with the supplied observations. This objective differs from conventional image-captioning or text-generation evaluation, for which similarity to a reference sentence can be a useful approximation of output quality. In radar interpretation, independently written expert accounts can describe the same physical evolution while emphasizing different, mutually compatible evidence. Conversely, a text can remain lexically similar to a reference after changing a scientifically decisive attribute such as motion direction, intensity trend, or temporal order. Reference-based NLP metrics therefore measure linguistic similarity rather than observation-grounded correctness.

C.1 An Inter-Expert Case Study

We illustrate this limitation using two independently authored, peer-reviewed analyses of the 20 July 2021 Zhengzhou extreme-rainfall event. The selected descriptions cover overlapping stages of the same convective development, including 1400–1600 Beijing Standard Time (BST). Expert A is based on Zhang et al. (2023, journal p. 433, Fig. 2e–f; 0600–0800 UTC), published in *Advances in Atmospheric Sciences* (doi:10.1007/s00376-022-2189-9). Expert B is based on Yin et al. (2022, journal pp. 11–12, Fig. 7a–c; 1130–1630 BST), published in the *Journal of Meteorological Research* (doi:10.1007/s13351-022-1166-7). Only the English descriptions used for comparison are presented below.

Expert A: English description.

> Note that a superior cloud cluster was present over southwestern Zhengzhou City, which was evidenced by a small area where stronger radar echoes exceeded 50 dBZ at 0600 UTC 20 July (Fig. 2e), leading to the more intense rainfall locally (Fig. 1c). Two hours later, the superior cloud cluster with further intensified radar echoes was located over Zhengzhou City by 0800 UTC 20 July (Fig. 2f), causing more extreme heavy rainfall to occur within the hour from 0800 to 0900 UTC, corresponding to the strongest hourly rainfall in Zhengzhou City (Fig. 1c). Note that the MβCV center also migrated slightly northwestward (Fig. 2f) as compared to the location two hours earlier (Fig. 2e), which was about 100 km away from the super cloud cluster over Zhengzhou City (Fig. 2f), indicating that a dynamical connection may have existed between the MβCV and super cloud cluster for the extreme hourly rainstorm over Zhengzhou City in terms of mesoscale dynamics (as discussed in Section 5).

Expert B: English description.

> In the observations (Fig. 7a), several convective cells showed up at 1130 BST 20 July along the east to north edge of the Songshan Mountain over 250–500-m altitudes. Meanwhile, an obvious coverage of trailing stratiform rainfall appeared on the north side of the dense equivalent potential temperature band (θ_e front) under the influence of the mid-tropospheric south-to-southwesterly flows. After 2.5 h, only two convective cells continued to develop, and others dissipated (Fig. 7b). Subsequently, the two convective cells merged around 1530 BST, and a well-organized convective cell formed over Zhengzhou at about 1630 BST (Fig. 7c). Once formed, the storm began its rapid development, with its peak radar reflectivity over 58 dBZ, attributable to the distinct northwestward expansion of the high-θ_e tongue (Figs. 7b, c).

The two selected accounts emphasize complementary aspects of the event rather than presenting directly contradictory interpretations. Expert A emphasizes the intensification and location of the strong echo cluster relative to the MβCV. Expert B emphasizes cell survival, merger, organization, peak reflectivity, and the high-θ_e tongue. Both accounts identify rapid convective intensification over Zhengzhou during overlapping development stages, but they select different physical details and use different discourse structures.

C.2 Metric Protocol

We treated each expert passage once as the candidate and once as the reference, and report the mean of the two directions. ROUGE-L is the F1 score based on the longest common subsequence with Porter stemming. CIDEr follows the COCO implementation with one- to four-gram features and $\sigma = 6$. Because CIDEr requires corpus-level inverse document frequency (IDF), we used 144 deduplicated radar- and convection-related sentences extracted from the two source papers as the background corpus. With only the target pair, the standard one-document CIDEr calculation returns exactly zero because its IDF estimate is degenerate; this zero must not be interpreted as scientific disagreement. BERTScore uses `roberta-large`, layer 17, without IDF. We report both its raw F1 and the official English baseline-rescaled F1. The full source passages contain 162 and 125 whitespace-delimited words, respectively.

Table 7: Text-similarity scores between two valid expert descriptions of the same Zhengzhou event and overlapping development stage. ROUGE-L and CIDEr are reported in their standard forms, while BERTScore is reported both as the raw F1 score and after the official baseline rescaling.

Metric	Symmetric score
ROUGE-L F1	0.1347
CIDEr	3.68×10^{-7}
BERTScore F1 (raw)	0.8560
BERTScore F1 (baseline-rescaled)	0.1467

C.3 Interpretation

ROUGE-L assigns a low score because the experts use different vocabulary, sentence organization, and selections of physical evidence. The corpus-backed CIDEr score is effectively zero because there is little shared higher-order n-gram content and because the metric is sensitive to both its IDF corpus and the difference in passage length. These low values do not indicate that either analysis is scientifically invalid. Instead, they show that lexical consensus is a poor proxy for agreement when two experts provide complementary meteorological interpretations.

Raw BERTScore gives a seemingly high F1 of 0.8560, whereas the official baseline-rescaled value is 0.1467. The substantial numerical difference reflects the change of scale introduced by baseline rescaling rather than a change in the underlying comparison. More importantly, neither form of BERTScore verifies whether specific claims about location, timing, motion, intensity, or storm organization are supported by the radar observations. A fluent but directionally incorrect interpretation could therefore remain close to a correct reference in embedding space.

This case demonstrates that ROUGE-L, CIDEr, and BERTScore answer a different question from the one posed by radar interpretation. They may be retained as auxiliary diagnostics of wording or broad semantic relatedness, but they cannot distinguish a complementary valid analysis from an unsupported radar claim. We consequently use radar-specific factual dimensions—motion, intensity, structure, evolution, and temporal consistency—together with domain-expert judgment as the primary evaluation criteria. This two-expert comparison is an illustrative case study rather than population-level statistical evidence; its purpose is to expose the construct mismatch between generic text similarity and observation-grounded scientific correctness.

D Implementation Details

We specialize Qwen3-VL-8B using LoRA-based supervised fine-tuning. Unless otherwise specified, LoRA adapters are inserted into both the self-attention projection layers (query, key, value, and output projections) and the MLP projection layers (gate, up, and down projections). The vision encoder and multimodal projector are kept frozen, and no additional radar-specific prediction head is introduced. All controlled supervision-quality comparisons use the same optimization configuration; only the source of the training supervision is changed.

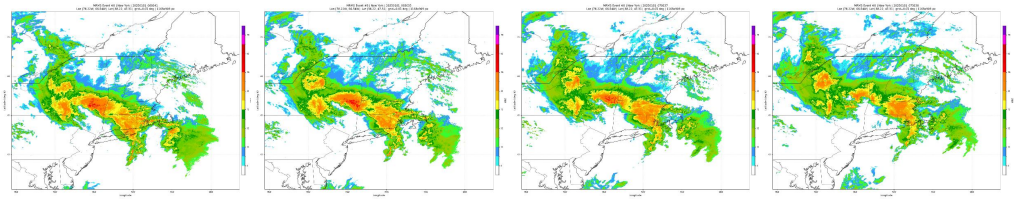
Training uses four A100 GPUs with a per-device batch size of one and four gradient-accumulation steps, giving an effective global batch size of 16. We use AdamW with a peak learning rate of 1×10^{-6} and cosine decay. Training is performed in BF16, with TF32 enabled.

For automatic supervision construction, Qwen3-VL-32B-Instruct is used for candidate generation, candidate-blind independent auditing, and cross-verification. Candidate generation uses temperature 0.20 and top-p 0.90 to retain limited generation diversity. Auditing and cross-verification use deterministic decoding with temperature 0.00 and top-p 1.00. The maximum output length is 16,384 tokens for all three stages. The model is served in BF16 using vLLM with four-way tensor parallelism across four A100 GPUs.

Table 8: Implementation details for fine-tuning and supervision construction.

Configuration	Setting
Base model	Qwen3-VL-8B
Fine-tuning method	LoRA-based supervised fine-tuning
LoRA target modules	Attention projections (q, k, v, and o) and MLP projections (gate, up, and down)
LoRA rank r	16
LoRA scaling factor α	32
LoRA dropout	0.05
Peak learning rate	1×10^{-6}
Per-device batch size	1 multimodal sample
Gradient accumulation	4 steps
Effective global batch size	$1 \times 4 \times 4 = 16$ multimodal samples
Optimizer	AdamW
Learning-rate scheduler	Cosine decay
Training precision	bfloat16 (BF16), with TF32 enabled
Hardware	$4\times$ NVIDIA A100 GPUs
Random seed	42 (Hugging Face default; not overridden)
Candidate-generation model	Qwen3-VL-32B-Instruct
Candidate-generation decoding	Temperature 0.20; top-p 0.90; maximum 16,384 output tokens
Independent-audit decoding	Temperature 0.00; top-p 1.00; maximum 16,384 output tokens
Cross-verification decoding	Temperature 0.00; top-p 1.00; maximum 16,384 output tokens
Generation serving	vLLM, BF16, four-way tensor parallelism on $4\times$ NVIDIA A100 GPUs

Prompt : *Act as a meteorological expert with strong experience in severe convective radar analysis and produce event-level evolution analysis from time-sorted radar composite reflectivity frames.*



Ground Truth: *At 06:00:41 UTC on 1 January 2025, a banded echo system stretching from western New York State to eastern Massachusetts appeared in the MRMS composite reflectivity imagery. The system exhibited a northwest-to-southeast orientation, with its primary intensity maintained at 30–40 dBZ. An intense core with reflectivity of 50–60 dBZ was present near 74.5 °W, indicating well-organized structure and active localized convection in the early stage of the system.The system then propagated steadily eastward. By 06:30:35 UTC, the western end of its major axis had shifted to around 77 °W, while the original intense core moved eastward to 73.5 °W and weakened to 45–50 dBZ, signalling the onset of core intensity decay as the system translated.At 07:00:37 UTC, the system continued its eastward motion. The former intense core further weakened to 40–45 dBZ, and a new intense core of 50–60 dBZ emerged near 72 °W, marking internal core reorganization and a shift of the dominant convective center within the system.By 07:30:38 UTC, the entire system had advanced into the region between 75 °W and 70 °W. The newly formed intense core retained an intensity of 50–60 dBZ, whereas the original core kept weakening and its peripheral regions tended to disperse. The overall organization of the system degraded compared with its initial phase.*

Ours: *At 06:00:41 on January 1, 2025, a banded echo system extending from western New York State to eastern Massachusetts became evident in the MRMS composite reflectivity imagery, with an overall northwest–southeast orientation. The main body maintained reflectivity values of 30–40 dBZ, while a strong core near 74.5 °W reached 50–60 dBZ, indicating relatively strong organization and locally active convection during the initial stage. The system subsequently continued to move eastward. By 06:30:35, the western end of the main axis had shifted to around 77 °W, while the original strong core had moved eastward to 73.5 °W and weakened to 45–50 dBZ, suggesting that the core intensity had begun to decay as the system progressed. At 07:00:37, the system continued its eastward movement, with the original strong core further weakening to 40–45 dBZ, while a new strong core with reflectivity of 50–60 dBZ emerged near 72 °W. This indicated internal core reorganization and a shift in the dominant convective center. By 07:30:38, the system had moved eastward to the 75 °W–70 °W region. The new strong core maintained an intensity of 50–60 dBZ, while the original core continued to weaken, and the peripheral regions became increasingly fragmented. Overall, the system exhibited a lower degree of organization than during the initial stage.*

Figure 3: Results Presentation

E Qualitative Radar-Event Descriptions and Scoring Examples

F Dimension-Level GPT-5.6 Sol Evaluation and Graphical Illustration of Partial Main-Text Results

Table 9: Dimension-level evaluation results obtained using GPT-5.6 Sol. All scores are percentages and higher is better. GPT Score is the arithmetic mean of Motion, Intensity, Structure, Evolution, and Temporal scores.

Experiment	Model / Setting	Scale	Motion	Intensity	Structure	Evolution	Temporal	GPT Score
MRMS Main	Qwen3-VL-8B	8B	58.0	64.0	61.0	59.5	62.5	61.0
	InternVL3.5-8B	8.5B	62.1	66.8	63.6	64.9	64.1	64.3
	MiniCPM-V 4.5	8B	63.8	68.7	66.4	64.9	66.7	66.1
	GLM-4.1V-9B-Thinking	9B	68.1	72.9	69.6	71.5	68.9	70.2
	Qwen3-VL-32B	32B	77.8	82.1	78.0	81.0	78.1	79.4
	RadarVL-8B	8B	**82.7**	80.4	**79.3**	**82.2**	**78.4**	**80.6**
Verification Ablation	Raw candidate annotation SFT	8B	65.9	69.8	66.5	68.7	66.1	67.4
	Audit + cross-verified supervision SFT	8B	78.4	74.9	73.1	77.8	73.8	75.6
LoRA Ablation	Attention only	8B	78.2	76.0	74.7	79.1	76.0	76.8
	MLP only	8B	73.9	78.4	76.8	74.1	72.3	75.1
	Attention + MLP	8B	**82.7**	**80.4**	**79.3**	**82.2**	**78.4**	**80.6**
OPERA Zero-shot	Qwen3-VL-8B	8B	55.7	62.8	59.4	56.9	59.7	58.9
	Qwen3-VL-32B	32B	70.3	**75.8**	74.1	73.6	69.7	72.7
	RadarVL-8B	8B	**77.5**	73.9	**76.1**	**75.4**	**70.1**	**74.6**

As described in Section 3.1, GPT-5.6 Sol is used as a supplementary automated evaluator under the same five radar-understanding dimensions and the same three-level scoring rubric used in the expert evaluation. Specifically, each model output is evaluated in terms of motion direction, echo intensity, structure, evolution trend, and temporal consistency.

Table 9 reports the complete dimension-level GPT-5.6 Sol evaluation results. Each dimension score is averaged over all evaluated events and reported as a percentage. The *GPT Score* reported in the main paper is the arithmetic mean of these five dimension-level scores. GPT-5.6 Sol evaluation is used only as a supplementary assessment; the domain-expert-based Radar Score remains the primary evaluation metric.

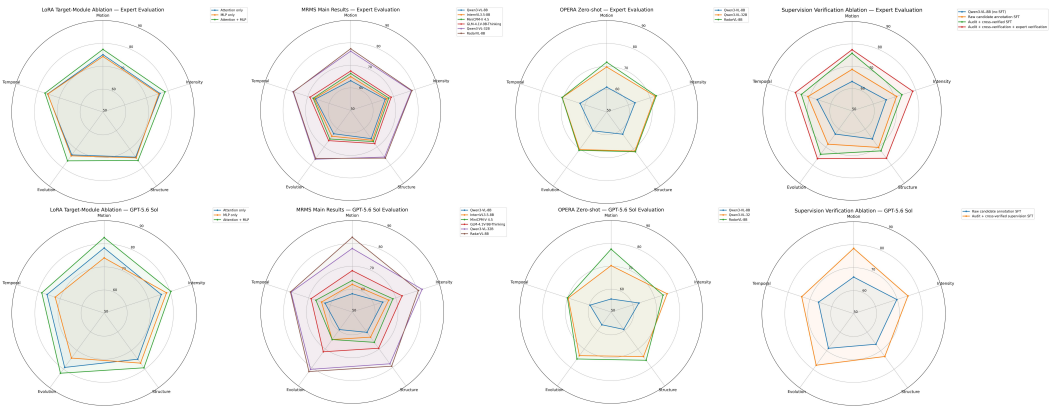


Figure 4: Radar chart presentation of partial results in the main text and appendices