

---

# RadarQA: Multi-modal Quality Analysis of Weather Radar Forecasts

---

Anonymous Author(s)

Affiliation

Address

email

## Appendix

### A Overview

This Appendix is structured as follows. Dataset details are described in Appendix B. More ablation studies, qualitative and quantitative results are presented in Appendix C

### B Dataset Details

#### B.1 Details of Scientific Attribute Library

To facilitate dataset construction, we design a scientific attribute library grounded in physical principles. This library comprises 5 super-categories and 10 sub-categories, comprising 35 attributes. Combined with the overall performance of the predictions at both the frame and sequence levels, these constitute a total of 37 key attributes used for dataset construction. The definitions of the 35 attributes in our scientific attribute library are provided in detail below.

##### Intensity.

- **Miss.** (a) Miss Performance. The proportion of regions with observed precipitation in the ground truth that are incorrectly predicted as “sunny” in the forecast. (b) Raw Rainfall Level. The rainfall levels in the ground truth for regions where rainfall is missed in the prediction. (c) Miss Rainfall Level. The rainfall levels in the prediction for regions where rainfall is missed. (d) Miss Direction. The directions in the prediction in which specific rainfall levels that are missed in the prediction.
- **FAR.** (a) FAR Performance. The proportion of regions labeled as “sunny” in the ground truth but incorrectly predicted with precipitation. (b) Raw Rainfall Level. The rainfall levels in the ground truth for regions where rainfall is falsely alarmed. (c) FAR Rainfall Level. The rainfall levels in the prediction for regions where rainfall is falsely alarmed. (d) FAR Direction. The directions in the prediction in which specific false-alarm rainfall levels that appear in the prediction.
- **High Value Construction.** (a) High Value Retain Performance. The ability of the prediction to consistently preserve high-value regions. (b) High Value Mismatch Type (Sequence). The type of mismatch in regions with high values (*i.e.*, precipitation at “intense” level or above) across the prediction and ground truth sequence. (c) High Value Mismatch Direction (Sequence). The directions in which high-value regions were mismatched. (d) High Value Mismatch Performance. The ability of the prediction to predict intense precipitation levels. (e) High Value Mismatch Type (Frame). The type of mismatch in regions with high values (*i.e.*, precipitation at “intense” level or above) across the prediction and ground truth frame. (f) High Value Mismatch Direction (Frame). The directions in which high-value regions were mismatched. (g) Max Rainfall Level. The maximum precipitation level in the observation.

##### Precipitation Conservation.

Table A1: **Characteristics** of each attribute in terms of level (frame / sequence), reference type (caption / comparison), annotation method (human / automation), and usage purpose (rating / assessment).

Attributes	Level		Reference		Annotation		Usage	
	Frame	Sequence	Caption	Comparison	Human	Automation	Rating	Assessment
Miss Performance	✓	✗	✗	✓	✗	✓	✓	✓
Raw Rainfall Level for Miss	✓	✗	✗	✓	✗	✓	✗	✓
Miss Rainfall Level	✓	✗	✗	✓	✗	✓	✗	✓
Miss Direction	✓	✗	✗	✓	✗	✓	✗	✓
FAR Performance	✓	✗	✗	✓	✗	✓	✓	✓
Raw Rainfall Level for FAR	✓	✗	✗	✓	✗	✓	✗	✓
FAR Rainfall Level	✓	✗	✗	✓	✗	✓	✗	✓
FAR Direction	✓	✗	✗	✓	✗	✓	✗	✓
High Value Retain Performance	✗	✓	✗	✓	✗	✓	✓	✓
High Value Mismatch Type (sequence)	✗	✓	✗	✓	✗	✓	✗	✓
High Value Mismatch Direction (sequence)	✗	✓	✗	✓	✗	✓	✗	✓
High Value Mismatch Performance	✓	✗	✗	✓	✗	✓	✓	✓
High Value Mismatch Type (Frame)	✓	✗	✗	✓	✗	✓	✗	✓
High Value Mismatch Direction (Frame)	✓	✗	✗	✓	✗	✓	✗	✓
Max Rainfall Level	✓	✗	✓	✗	✗	✓	✗	✓
Cumulate Precipitation Performance	✗	✓	✗	✓	✗	✓	✓	✓
Cumulate Precipitation Difference	✗	✓	✗	✓	✗	✓	✗	✓
Mismatch Direction	✗	✓	✗	✓	✗	✓	✗	✓
Shape Change	✗	✓	✓	✗	✓	✗	✗	✓
Scale Change	✗	✓	✓	✗	✓	✗	✗	✓
Convective Cell Change	✗	✓	✓	✗	✓	✗	✗	✓
Intensity Change	✗	✓	✓	✗	✓	✗	✗	✓
Dynamic Consistency Performance	✗	✓	✗	✓	✓	✗	✓	✓
Move Direction	✗	✓	✓	✗	✓	✗	✗	✓
Speed Difference	✗	✓	✗	✓	✓	✗	✗	✓
Rotation Center	✗	✓	✓	✗	✓	✗	✗	✓
Difference in Generation	✗	✓	✗	✓	✓	✗	✗	✓
Difference in Dissipation	✗	✓	✗	✓	✓	✗	✗	✓
Sharpness Performance	✓	✗	✗	✓	✗	✓	✓	✓
Shape Type	✗	✓	✓	✗	✓	✗	✗	✓
Shape Mismatch Direction	✗	✓	✗	✓	✓	✗	✗	✓
Shape Mismatch Reason	✗	✓	✗	✓	✓	✗	✗	✓
Artifacts Direction	✗	✓	✗	✓	✓	✗	✗	✓
Organization Degree	✗	✓	✓	✗	✓	✗	✗	✓
Distribution	✓	✗	✓	✗	✗	✓	✗	✓
Overall Performance (Sequence)	✗	✓	✗	✓	✓	✗	✓	✓
Overall Performance (Frame)	✓	✗	✗	✓	✓	✗	✓	✓

- **Cumulate Precipitation.** (a) Cumulate Precipitation Performance. The degree to which the cumulative precipitation predicted over the entire sequence aligns with the ground truth. (b) Cumulate Precipitation Difference. Differences between the total precipitation of the prediction and the ground truth across the sequence, indicating whether the forecast overestimates or underestimates cumulative rainfall. (c) Mismatch Direction. The directions in which the prediction fails to reconstruct the cumulative precipitation accurately.

#### Precipitation Dynamic Distribution.

- **Morphogenesis.** (a) Shape Change. The change in the shape of the convective system over time in the ground truth. (b) Scale Change. The change in the spatial area of the convective system across frames in the ground truth. (c) Convective Cell Change. The change in the number of convective cells. (d) Intensity Change. The change in the precipitation intensity over time. (e) Dynamic Consistency Performance. The overall consistency of dynamic evolution between the prediction and the ground truth.
- **Trajectory.** (a) Move Direction. The primary direction of movement of the convective system in the ground truth. (b) Speed Difference. The difference in the movement speed of the convective system between the prediction and the ground truth. (c) Rotation Center. The spatial location that acts as the center of rotation for convective system evolution.

#### Convective Cycle.

Table A2: **Statistics** of RawRQA-20K.

Event type	Flash flood	Flood	Funnel cloud	Hail	Heavy rain	Thunderstorm wind	Tornado
# of events	218	121	58	556	55	1030	121

52 • **Genesis.** (a) Difference in Generation. The difference in the number of newly generated  
53 convective cells between the prediction and the ground truth over the entire sequence.

54 • **Dissipation.** (a) Difference in Dissipation. The difference in the number of dissipated convective  
55 cells between the prediction and the ground truth throughout the sequence.

## 56 **Morphology.**

57 • **Sharpness.** (a) Sharpness Performance. The degree of similarity between the fine-grained  
58 contours in the prediction and those in the ground truth.

59 • **Shape.** (a) Shape Type. The morphological pattern of the convective system in the observation.  
60 (b) Shape Mismatch Direction. The directions in which the evolution trend of the convective  
61 shape in the prediction diverges from that in the ground truth. (c) Shape Mismatch Reason. The  
62 underlying cause contributing to the mismatch in convective morphology between the prediction  
63 and observation. (d) Artifacts Direction. The directions in which artificial patterns appear in the  
64 predicted sequence that do not exist in the observation. (e) Organization Degree. The temporal  
65 trend of structural organization in the ground truth reflects how orderly the convective system is  
66 over time. (f) Distribution. The directional distribution of precipitation in the observation

67 An overview of the properties associated with each attribute is demonstrated in Tab. A1.

## 68 **B.2 Details of Raw Data Statistics**

69 To ensure the diversity of samples in RawRQA-20K, we consider both a wide range of storm event  
70 types and a diverse set of generative models. First, our RawRQA-20K covers seven storm event types,  
71 including flash flood, flood, funnel cloud, hail, heavy rain, thunderstorm wind, and tornado. Due to  
72 their strong convective nature and high impact, these storm events pose significant challenges for  
73 forecasting and contribute to a diverse sample space. The number of samples for each event type is  
74 summarized in Tab. A2. The coverage area of storm events is shown in Fig. A2.

75 We employ a total of seven representative nowcasting models to generate prediction samples. As  
76 illustrated in Fig. A1, these models produce diverse samples that reflect a wide range of forecast  
77 qualities. For example, Cascast tends to over-predict in high-value regions, yet generally exhibits  
78 superior performance in detail reconstruction and dynamic consistency. In contrast, DGMN often  
79 introduces substantial artifacts, which significantly degrade the overall quality. Meanwhile, PredRNN  
80 suffers from severe temporal blurring and exhibits poor performance in “high value retain”. These  
81 varied quality issues are reflected in the corresponding differences across the assessment reports.

## 82 **B.3 Details of Human Annotation Questionnaire**

83 For the human-annotated attributes listed in Tab. A1, we employed an annotation pipeline to ensure  
84 consistency and quality. First, for each attribute, we designed a corresponding multiple-choice  
85 question, with domain experts defining clear annotation guidelines. Second, a small set of pilot  
86 samples was used to evaluate annotation quality from several annotation companies. The company  
87 with the most accurate performance was selected for large-scale annotation. Third, all annotators  
88 underwent standardized training to align their understanding with expert standards. Each annotator  
89 completed a trial annotation set, which was reviewed by experts who provided feedback and corrected  
90 any misinterpretations. Fourth, upon completion of each annotation batch, a cross-validation step  
91 is conducted by different annotators to ensure quality. Finally, after annotation, domain experts  
92 performed quality control by randomly sampling and reviewing 35% of the samples in each batch. A  
93 batch would be accepted only if the sampled annotations met the quality standards; otherwise, the  
94 annotators were required to re-annotate the entire batch.

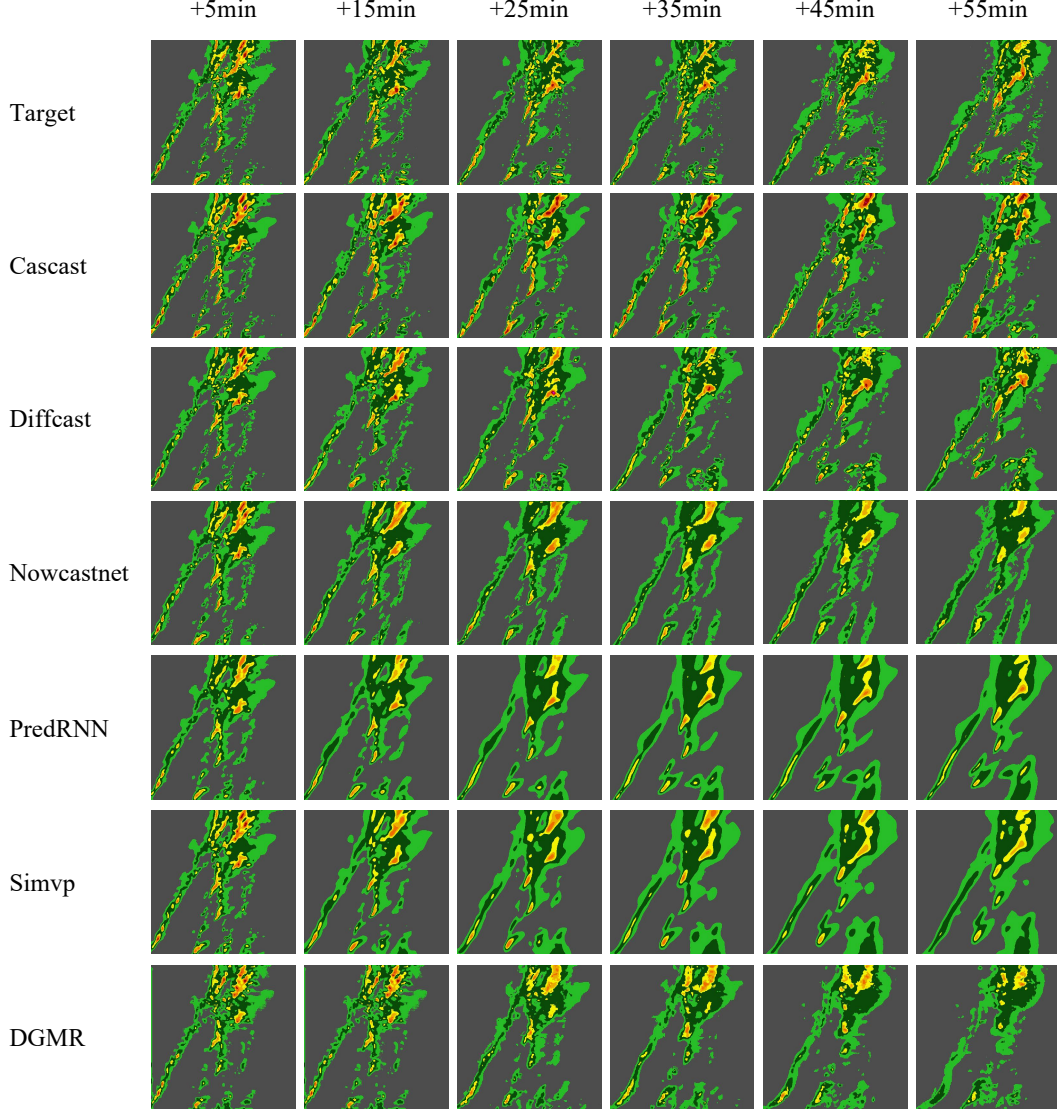


Figure A1: A set of example forecasts on SEVIR.

#### 95 B.4 Automated Generation

96 As shown in Tab. A1, 20 attributes are grounded in score-based metrics, where automated annotation  
 97 provides more precise and consistent results compared to manual labeling. In this process, all  
 98 the required thresholds or parameters are determined with the assistance of domain experts. The  
 99 corresponding computation procedures for these attributes are detailed below.

#### 100 General Attributes.

- 101 • False Alarm Performance. First, we calculate the false alarm rate. Let  $\mathcal{G}$  and  $\mathcal{P}$  denote the sets  
 102 of pixels with precipitation in the ground truth and the prediction, respectively. Define Hits as  
 103  $H = |\mathcal{G} \cap \mathcal{P}|$  and False Alarms as  $(F = |\mathcal{P} \setminus \mathcal{G}|)$ . The false alarm rate is given by:

$$\text{false alarm rate} = \frac{F}{H + F} \quad (\text{A1})$$

104 Thresholds [0.1, 0.2, 0.3] are selected to categorize the false alarm rate into four performance  
 105 levels (“Great”, “Good”, “Fair”, “Poor”).

- 106 • Miss Performance. Similar to the false alarm rate, we compute the miss rate based on the binary  
 107 masks. Following SEVIR, we define a pixel as having precipitation if its value exceeds 16. Let  $\mathcal{G}$

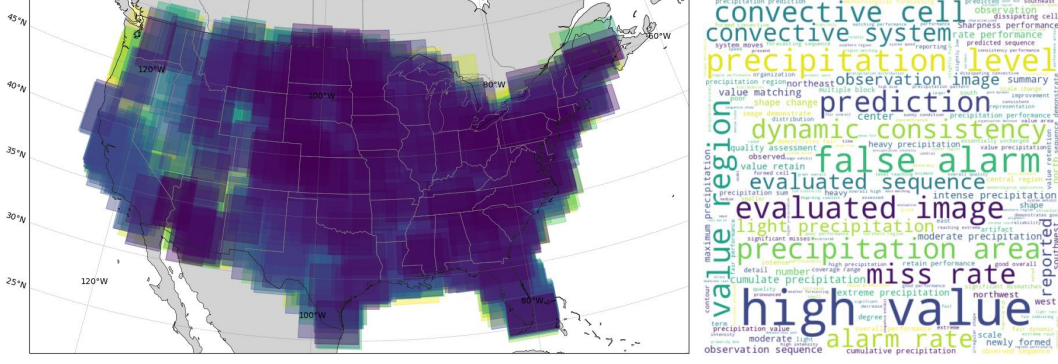


Figure A2: **Coverage area** of selected storm events in our RQA-70K dataset, which spans across the CONUS region. Figure A3: **Wordcloud map** of our introduced RQA-70K dataset.

Caption		
1. What is the moving direction of the convective system? A. ↑ B. ↓ C. → D. ← E. ↘ F. ↙ G. ↗ H. ↖ I. ∅ J. ∅	4. What is the rotate center of the convective system? A. ↑ B. ↓ C. → D. ← E. ↘ F. ↙ G. ↗ H. ↖ I. center J. no rotation	7. What is the shape of convective system? A. scattered F. multi-block-like B. banded G. multi-arc-shaped C. block-like H. multi-banded D. large patch-like I. spiral shaped E. arc shaped J. Irregular shaped
2. How does the number of convective cells change? A. increase C. remain the same B. decrease	5. How does the coverage area of convective system change? A. increase C. remain the same B. decrease	8. How does the shape of convective system change? A. merge E. split B. stretch F. disappear C. shrink G. form D. dilate H. remain the same
3. How does the intensity of convective system change? A. increase C. remain the same B. decrease	6. How does the organization degree of convective system change? A. increase C. remain the same B. decrease	

Comparison		
1. In which directions are the diff. in shape change most severe? A. ↑ B. ↓ C. → D. ← E. ↘ F. ↙ G. ↗ H. ↖ I. center J. remains the same	3. What are the directions that have artifacts? A. ↑ B. ↓ C. → D. ← E. ↘ F. ↙ G. ↗ H. ↖ I. center J. remains the same	5. The scale of dissipated convective cell in the prediction is A. larger C. basically the same B. smaller
2. What is the main issue within the direct. with most diff. in shape change? A. scale diff. C. position diff. B. diff. of convective cell numbers	4. The scale of generated convective cell in the prediction is A. larger C. basically the same B. smaller	6. The movement speed of the convective cycle in the prediction is A. faster C. basically the same B. slower

Rating		
1. What is the overall performance of the predicted sequence? A. great B. good C. fair D. poor	2. What is the dynamic consistency performance of the predicted sequence? A. great B. good C. fair D. poor	3. What is the overall performance of the predicted image? A. great B. good C. fair D. poor

Figure A4: **Human annotation questionnaire** for the 17 attributes that require manual labeling.

108 and  $\mathcal{P}$  denote the sets of pixels with precipitation in the ground truth and prediction. Define Hits  
109 as  $H = |\mathcal{G} \cap \mathcal{P}|$  and Misses as  $M = |\mathcal{G} \setminus \mathcal{P}|$ . The miss rate is defined as:

$$\text{miss rate} = \frac{M}{H + M} \quad (\text{A2})$$

110 Thresholds[0.1, 0.2, 0.4] are used to categorize the miss rate into four performance levels.

111 • **Sharpness Performance.** Following SRViT, we evaluate the sharpness of the prediction and the  
112 ground truth using the Sobel filter. Specifically, let  $S_{gt}$  and  $S_{pred}$  denote the mean Sobel value of  
113 the ground truth and the prediction, respectively:

$$S_{gt} = \frac{1}{N} \sum_{i=1}^n \text{Sobel}(\text{gt})_i, S_{pred} = \frac{1}{N} \sum_{i=1}^n \text{Sobel}(\text{Pred})_i \quad (\text{A3})$$



Table A3: Structure of **detailed descriptions** for each general attribute.

General Attributes	Detailed Description
High Value Mismatch	In the <i>high value mismatch direction</i> , the prediction is <i>high value mismatch type</i> (over-predict / under-predict).
Miss	In the <i>Miss direction</i> , the <i>raw rainfall level</i> is misclassified as <i>miss rainfall level</i> .
Cumulate Precipitation	In the <i>mismatch direction</i> , the cumulate precipitation is <i>cumulate precipitation difference</i> .
High Value Retain	In the <i>high value mismatch direction</i> , the prediction is <i>high value mismatch type</i> (over-predict / under-predict).

We then compute the relative difference:

$$d = \begin{cases} 2 - \left| \frac{S_{pred}}{S_{gt}} \right|, & \text{if } \left| \frac{S_{pred}}{S_{gt}} \right| > 1 \\ \left| \frac{S_{pred}}{S_{gt}} \right|, & \text{otherwise} \end{cases} \quad (A4)$$

Finally, we clip negative values to zero, and define the sharpness score as:

$$\text{sharpness score} = \max(0, d) \quad (A5)$$

Thresholds [0.5, 0.7, 0.9] are used to categorize the sharpness into four levels.

- **High Value Mismatch Performance.** We first count the number of high-value pixels in both the prediction and the ground truth (*i.e.*, pixels with intensity values greater than 219), denoted as  $N_{pred}$  and  $N_{gt}$ , respectively. The relative error is computed as:

$$\mathcal{E}_{rel} = \left| \frac{N_{gt} - N_{pred}}{N_{gt}} \right| \quad (A6)$$

The high value mismatch score is subsequently defined as

$$\text{high value mismatch score} = \min(1, \max(0, 1 - \mathcal{E})) \quad (A7)$$

Thresholds [0.3, 0.6, 0.8] are used to categorize the high value mismatch into four levels.

- **High Value Retain Performance.** The high-value retain score is computed as the average high-value mismatch score across all frames. The same thresholds [0.3, 0.6, 0.8] are used to categorize the performance into four levels.
- **Cumulate Precipitation Performance.** First, we compute the total precipitation in the prediction and ground truth, denoted as  $P_{pred}$  and  $P_{gt}$ , respectively. We then calculate the relative precipitation error and define the cumulate precipitation score using the same method as in the computation of  $\mathcal{E}_{rel}$  and the high-value mismatch score. Thresholds [0.93, 0.97, 0.99] are applied to categorize the performance levels.

To provide a detailed characterization of the general attributes, we divide each image into a  $3 \times 3$  grid, resulting in nine spatial regions corresponding to nine directional sectors. For each general attribute, its detailed description is formulated as a combination of directional information and the associated prediction issue. For example, in the case of false alarms, a typical description takes the form of “in the *FAR direction*, the *raw rainfall level* is false alarmed as the *FAR rainfall level*.” This expression involves three distinct attributes, whose construction is detailed below.

**Raw Rainfall Level.** First, we compute the number of missed pixels for each rainfall intensity level. To incorporate the varying importance of different rainfall levels, we align with domain experts and assign weights [1, 1.5, 2.5, 5, 10, 20], corresponding to increasing rainfall intensity from “light” to “extreme”. Higher rainfall levels are given greater emphasis. We then compute the weighted sum of missed pixels for each level, ranking them in descending order, and identify the rainfall level with the highest weighted missing pixel count.

**FAR Rainfall Level.** For each raw rainfall level, we examine the corresponding locations in the prediction and count the occurrences of each predicted rainfall level. The rainfall level with the highest pixel count that is lighter than the raw rainfall level is selected as the FAR rainfall level.

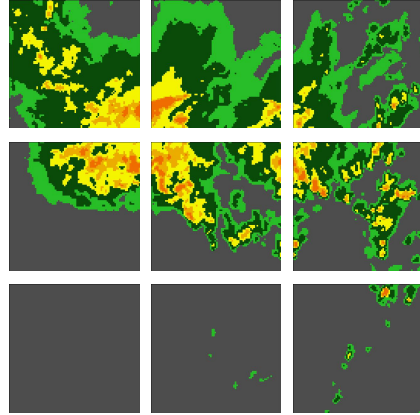


Figure A5: **Gridding** of the image into  $3 \times 3$  patches, each representing a directional sector.

Table A4: More results on ablation studies of multi-stage training strategy on **rating tasks**. Accuracy is used as the metric for the frame rating task.

Stage-1	Stage-2	Stage-3	Frame					Sequence			
			Overall	False Alarm	Miss	High Value Mismatch	Sharpness	Overall	Dynamic Consistency	Cumulate Precipitation	High Value Retain
✗	✗	✗	20.10	36.40	30.00	16.51	35.93	7.99	16.10	17.49	23.22
✓	✗	✗	60.93	63.37	61.63	63.02	71.28	61.42	42.44	42.20	74.53
✓	✓	✗	59.77	<b>68.14</b>	<b>67.67</b>	65.00	74.19	61.55	<b>64.17</b>	42.82	77.78
✓	✗	✓	61.28	65.00	66.40	<b>69.88</b>	<u>78.14</u>	<u>65.42</u>	52.31	<b>49.44</b>	<b>81.52</b>
✓	✓	✓	<b>61.51</b>	<u>65.35</u>	<b>67.67</b>	<u>69.19</u>	<b>78.60</b>	<b>66.17</b>	<u>53.31</u>	48.94	<u>80.52</u>

Table A5: More results on ablation studies of multi-stage training strategy on **assessment tasks**.

Stage-1	Stage-2	Stage-3	Frame					Sequence				
			BLEU	BERTScore	ROUGE_L	METEOR	GPT-4 Score	BLEU	BERTScore	ROUGE_L	METEOR	GPT-4 Score
✗	✗	✗	0.122	0.75	0.389	0.332	3.81	0.09	0.745	0.281	0.342	3.92
✓	✗	✗	0.195	0.799	0.498	0.417	6.40	<b>0.212</b>	0.812	0.429	0.453	6.22
✓	✗	✓	<u>0.212</u>	<b>0.810</b>	<u>0.511</u>	<b>0.423</b>	<u>6.83</u>	0.211	<b>0.816</b>	<u>0.431</u>	<b>0.461</b>	<u>6.56</u>
✓	✓	✓	<b>0.213</b>	<u>0.809</u>	<b>0.512</b>	<u>0.420</u>	<b>6.87</b>	<b>0.212</b>	<u>0.815</u>	<b>0.436</b>	<b>0.461</b>	<b>6.58</b>

152 *FAR Direction*. For each raw rainfall level, we compute the false alarm rate across different directions.  
153 We also count the number of pixels with raw rainfall level in each direction. To ensure both a high  
154 false alarm rate and a large false alarm area, We sort the directions by false alarm rate in descending  
155 order, and restrict our selection to those whose raw rainfall level pixel counts are among the top two.  
156 The first direction satisfying this condition is selected as the FAR direction.

157 For other general attributes, the structure of their detailed descriptions is summarized in Tab. A3, and  
158 the construction of their underlying attributes follows a similar procedure as in FAR.

## 159 C More Results

160 **Few-shot evaluation on frame rating task and frame assessment task**. We further evaluate  
161 the performance of different API-based models. As shown in Fig. A6, although other models are  
162 evaluated under few-shot settings, RadarQA consistently outperform all baselines without requiring  
163 any additional examples, demonstrating the effectiveness of RadarQA.

164 **Ablation studies on multi-stage training strategy**. For our multi-stage training strategy, we further  
165 examine the effectiveness of each stage across different metrics, as shown in Tab. A6 and Tab. A5.  
166 First, applying reinforcement learning significantly improves performance on reasoning-related  
167 metrics such as false alarm and miss rates. After supervised fine-tuning, the model leverages its ability  
168 on interpreting learned from assessment tasks as the reasoning step to better rate general attributes.  
169 Second, the full training strategy achieves the best performance on most metrics, demonstrating the  
170 effectiveness of the pipeline.

171 **Qualitative results**. More qualitative results of assessment tasks are shown in Fig. A7 and Fig. A8.

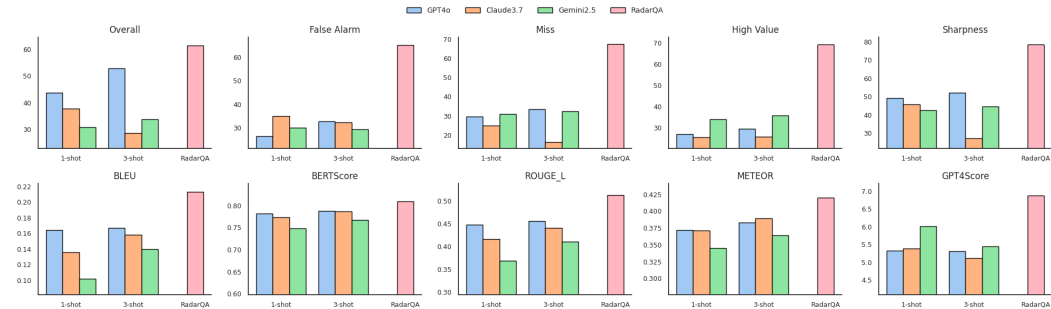


Figure A6: **Few-shot Results** on general attributes for the frame rating and frame assessment tasks. Accuracy is used as the metric for the frame rating task. RadarQA surpasses all methods.



Figure A7: Qualitative results on sequence assessment task.





Figure A8: Qualitative results on frame assessment task.

Table A6: Question pool of *frame rating* task..

#	Question
1	Could you score the predicted image based on miss, false alarm, sharpness, and high value matching, and then provide an overall performance level?
2	Please assign levels to the predicted image based on the four dimensions: miss, false alarm, sharpness, and high value matching, and give an overall performance level.
3	How would you score the quality of the evaluated image on the dimensions of miss, false alarm, sharpness, and high value matching, and what would the overall level be?
4	Can you score the predicted image using the four criteria: miss, false alarm, sharpness, and high value matching, and then provide an overall level?
5	Could you evaluate and score the predicted image using miss, false alarm, sharpness, and high value matching, then provide a final overall performance level?
6	How would you score the predicted image across dimensions of miss, false alarm, sharpness, and high value matching, and what would be the overall score?
7	Please score the evaluated image based on miss, false alarm, sharpness, and high value matching, then provide the overall performance level.
8	Could you score the image on miss, false alarm, sharpness, and high value matching, and then give an overall evaluation score for the image?
9	How would you rate the predicted image across the four dimensions, miss, false alarm, sharpness, and high value matching, and what is the overall performance level?
10	How would you rate the image on the four dimensions, miss, false alarm, sharpness and high value matching, and provide an overall performance level?

Table A7: Question pool of *frame assessment* task.

#	Question
1	Please start by describing the content of the observation image, and then evaluate the quality of the evaluated image based on dimensions miss_false alarm, sharpness and high-value-matching. Provide a comprehensive quality assessment report based on the 2 subtasks with summary.
2	How would you describe the observation image? Following that, could you evaluate the quality of the evaluated image across dimensions miss_false alarm, sharpness and high-value-matching, then give a summary?
3	Provide a detailed quality report of the evaluated image. First describe the content of the observation, then focus on miss, false alarm, sharpness, and high-value matching performance of predicted image.
4	Could you describe the observation image's content, then assess the quality of the evaluated image according to dimensions miss_false alarm, sharpness and high-value-matching in the format of a detailed report with summary?
5	Give a report of the evaluated image. First describe the content of the observation image, then focus on miss, false alarm, sharpness, and high-value matching performance of predicted image. Finally, summarize your analysis.
6	Please describe the observation image's content. Then, how would you assess the quality of the evaluated image based on miss_false alarm, sharpness and high-value-matching? Give a detailed report with summary.
7	What is your description of the observation image? Afterward, could you evaluate the quality of the evaluated image on miss_false alarm, sharpness and high-value-matching? Please provide a detailed report with summary.
8	Start by describing the content of the observation image, then assess the evaluated image on dimensions miss_false alarm, sharpness and high-value-matching. Provide a detailed report with summary.
9	How would you describe the content of the observation image? Then, how would you evaluate the quality of the evaluated image on miss_false alarm, sharpness and high-value-matching, and summarize your findings? Give a detailed report with summary.
10	What content description would you give for the observation image? Then, how would you evaluate the quality of the evaluated image across miss_false alarm, sharpness and high-value-matching. Provide a detailed final report with summary.

Table A8: Question pool of *sequence rating* task.

#	Question
1	Could you score the predicted sequence based on dynamic consistency, high value retaining and cumulate precipitation error, and then provide an overall performance level?
2	Please assign levels to the predicted sequence based on the four dimensions: dynamic consistency, high value retaining and cumulate precipitation error, and give an overall performance level.
3	How would you score the quality of evaluated image on the dimensions dynamic consistency, high value retaining and cumulate precipitation error, and what would the overall level be?
4	Can you score the predicted sequence using the four criteria: dynamic consistency, high value retaining and cumulate precipitation error, and then provide an overall level?
5	Could you evaluate and score the predicted sequence using dynamic consistency, high value retaining and cumulate precipitation error, then provide a final overall performance level?
6	How would you score the predicted sequence across dimensions dynamic consistency, high value retaining and cumulate precipitation error, and what would be the overall score?
7	Please score the evaluated sequence based on dynamic consistency, high value retaining and cumulate precipitation error, then provide the overall performance level.
8	Could you score the sequence on dynamic consistency, high value retaining and cumulate precipitation error, and then give an overall evaluation score for the sequence?
9	How would you rate the predicted sequence across the four dimensions, dynamic consistency, high value retaining and cumulate precipitation error, and what is the overall performance level?
10	How would you rate the sequence on the four dimensions, dynamic consistency, high value retaining and cumulate precipitation error, and provide an overall performance level?

Table A9: Question pool of *sequence assessment* task.

#	Question
1	Please start by describing the content of the observation sequence, and then evaluate the quality of the evaluated sequence based on dimensions of dynamic consistency, high value retain, and cumulate precipitation similarity. Provide a comprehensive quality assessment report based on the 2 subtasks with a summary.
2	How would you describe the observation sequence? Following that, could you evaluate the quality of the evaluated sequence across dimensions of dynamic consistency, high value retain, and cumulate precipitation similarity, then give a summary?
3	Provide a detailed quality report of the evaluated sequence. First describe the content of the observation, then focus on miss, false alarm, sharpness, and high-value matching performance of the predicted sequence.
4	Could you describe the observation sequence's content, then assess the quality of the evaluated sequence according to dimensions of dynamic consistency, high value retain, and cumulate precipitation similarity in the format of a detailed report with a summary?
5	Give a report of the evaluated sequence. First, describe the content of the observation sequence, then focus on miss, false alarm, sharpness, and high-value matching performance of predicted sequence. Finally, summarize your analysis.
6	Please describe the observation sequence's content. Then, how would you assess the quality of the evaluated sequence based on dynamic consistency, high value retain, and cumulate precipitation similarity? Give a detailed report with a summary.
7	What is your description of the observation sequence? Afterward, could you evaluate the quality of the evaluated sequence on dynamic consistency, high value retain, and cumulate precipitation similarity? Please provide a detailed report with a summary.
8	Start by describing the content of the observation sequence, then assess the evaluated sequence on dimensions of dynamic consistency, high value retain, and cumulate precipitation similarity. Provide a detailed report with a summary.
9	How would you describe the content of the observation sequence? Then, how would you evaluate the quality of the evaluated sequence on dynamic consistency, high value retain, and cumulate precipitation similarity, and summarize your findings? Give a detailed report with a summary.
10	What content description would you give for the observation sequence? Then, how would you evaluate the quality of the evaluated sequence across dynamic consistency, high value retain, and cumulate precipitation similarity? Provide a detailed final report with a summary.