

WildFireVQA: A Large-Scale Radiometric Thermal VQA Benchmark for Aerial Wildfire Monitoring

Supplementary Material

Supplementary Overview

This supplementary material complements the main paper by providing a detailed analysis of temperature-grounded retrieval, the complete WildFireVQA question inventory, and additional information on the multimodal inputs and annotation prompt structure used in the evaluation setting.

1. Fine-Grained Analysis of Temperature-Grounded Retrieval

To better understand when retrieved radiometric thermal statistics are beneficial, we perform a fine-grained analysis for Qwen3-VL [1] under the RGB setting by comparing the same model with and without RAG. We partition the 34 WildFireVQA questions into two groups: a temperature-related subset containing PD1, CL1, CMR4, DS3, DS7, and DS8, and a remaining-question subset containing all other questions. The grouped results in Table 1 show that the benefit of RAG is highly concentrated on the temperature-related questions. In particular, accuracy on this subset increases from 61.54% to 71.61%, yielding a gain of +10.07 points, while the remaining questions show essentially no improvement, changing from 51.51% to 51.15%.

This pattern indicates that retrieved radiometric thermal statistics are most useful when the question is directly aligned with the sensor-derived evidence appended to the prompt. In our setting, RAG provides a compact thermal summary from the paired radiometric TIFF, including temperature statistics and the fraction of pixels above physically meaningful thresholds such as 200°C and 400°C. The per-question results in Table 2 further support this interpretation. In particular, CMR4 benefits substantially because it requires reasoning about the hottest part of the fire, which is strongly supported by explicit temperature cues. CL1 also shows a clear improvement, suggesting that retrieved thermal evidence helps the model better distinguish active fire behavior from weaker or less intense burning conditions. Taken together, these results indicate that the proposed retrieval strategy is most effective when the model must reason about fire intensity and thermally grounded scene state, rather than relying only on visual appearance or broader scene interpretation alone.

2. Complete Question Set and Answer Space

Table 3 lists the complete WildFireVQA question set together with the canonical answer space for each question. The questions are grouped by the six task categories used

in the main paper: Presence and Detection, Classification, Distribution and Segmentation, Localization and Direction, Cross-Modal Reasoning, and Flight Planning. Defining a fixed answer space for each question helps reduce ambiguity and supports consistent comparison across models and evaluation settings.

3. Multimodal Inputs and Their Roles

As described in the main paper, each WildFireVQA sample contains three aligned modalities: an RGB image, a color-mapped thermal visualization, and a radiometric thermal TIFF. The RGB image provides visible-spectrum scene context, while the color-mapped thermal visualization provides a human-readable view of thermal structure. The paired radiometric thermal TIFF further provides per-pixel temperature values, enabling temperature-grounded reasoning and stronger validation of fire-related queries.

These modalities play different roles during dataset construction and benchmark evaluation. During dataset construction, answer generation uses the RGB image, the aligned color-mapped thermal visualization, and a compact radiometric summary derived from the paired thermal TIFF. During benchmark evaluation, the main paper studies controlled settings that isolate the effects of input modality and retrieval augmentation. More specifically, models are evaluated under RGB, Thermal, and corresponding retrieval-augmented variants in which compact radiometric thermal statistics are appended as auxiliary context.

This distinction is important for interpreting the benchmark. The annotation pipeline uses richer multimodal context to improve label reliability, whereas the evaluation protocol is designed to analyze how different MLLMs use RGB appearance, thermal structure, and retrieved numerical thermal cues when answering wildfire monitoring questions.

4. Annotation Protocol

For dataset-wide answer generation, the prompt follows the structure described in the main paper. Each prompt contains the RGB image, the aligned color-mapped thermal visualization, the question text, the candidate answer options, and a compact radiometric summary derived from the paired thermal TIFF. The radiometric summary includes the minimum temperature, maximum temperature, temperature standard deviation, percentage of pixels exceeding 200°C, and percentage of pixels exceeding 400°C. These quantities provide compact, physically grounded thermal context

Table 1. Fine-grained comparison of Qwen3-VL [1] under the RGB setting, with and without retrieved radiometric thermal statistics. We partition the 34 WildFireVQA questions into a temperature-related subset and a remaining-question subset. The temperature-related subset includes questions whose answers depend directly on hotspot presence, fire intensity, or temperature-threshold coverage. Gains are computed as $RAG - No-RAG$.

Group	Question IDs	Total	No-RAG Correct	No-RAG Acc.	RAG Correct	RAG Acc.	Gain
Temperature-related	PD1, CL1, CMR4, DS3, DS7, DS8	36,582	22,513	61.54	26,197	71.61	+10.07
Remaining questions	All others	170,716	87,944	51.51	87,328	51.15	-0.36

Table 2. Per-question analysis for for the temperature-related questions (CL1 and CMR4) that show positive gains under retrieved radiometric thermal statistics. Each question contains 6,097 instances. Gains are computed as $RAG - No-RAG$.

ID	Question	No-RAG Acc.	RAG Acc.	Gain
CL1	What is the dominant fire behavior observed in the scene?	48.93	67.69	+18.76
CMR4	What is the temperature of the hottest part of the fire in this scene in degrees Celsius?	58.88	88.60	+29.72

in addition to the visual evidence. A simplified form of the annotation prompt is shown below.

Prompt Template

You are provided with two aligned images of the same wildfire scene:

1. The first image is a standard RGB aerial image.
2. The second image is a color-mapped thermal image derived from radiometric thermal data.

Use both images together to understand wildfire activity in the scene.

You are also given a compact temperature summary computed from the paired radiometric thermal TIFF:

- Minimum temperature: $\{\min\}$
- Maximum temperature: $\{\max\}$
- Temperature standard deviation: $\{\text{std}\}$
- Percentage of pixels above 200°C: $\{\text{pct}_{200}\}$
- Percentage of pixels above 400°C: $\{\text{pct}_{400}\}$

Use the visual evidence and the temperature summary jointly, then answer the multiple-choice question by selecting the correct option.

This prompt structure follows the dataset-wide answer generation procedure described in the main paper and preserves a consistent multimodal labeling protocol across all samples and question types.

4.1. Example Annotation Prompt

Figure 1 shows an illustrative prompt following the format used during WildFireVQA dataset construction. The prompt combines aligned RGB and thermal inputs with a compact radiometric summary derived from the paired

thermal TIFF, then presents the question in multiple-choice form. This standardized structure is used to support consistent dataset-wide answer generation across the benchmark.

References

- [1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 2

Table 3. Comprehensive list of questions and answer choices grouped by category for wildfire analysis.

ID	Question	Answer Choices	ID	Question	Answer Choices
Presence/Detection			Distribution/Segmentation		
PD1	Are active thermal hotspots detected?	Yes / No	DS1	What is the spatial distribution of the active hotspots?	Scattered / Concentrated / Linear / No active hotspots
PD2	Is smoke visible?	Yes / No	DS2	How continuous is the fuel bed in the fire's potential path?	Continuous / Patchy / Discontinuous
PD3	Are visible flames present?	Yes / No	DS3	How consistent is the intensity of the active thermal hotspots in the scene?	Similar intensity / Different intensity / No active hotspots
PD4	Are any buildings or residential structures visible?	Yes / No	DS4	What proportion of visible vegetation is affected by fire, either actively burning or already burned?	1-25% / 25-50% / >50% / None
PD5	Are natural fuel breaks like rock outcroppings or sparse vegetation visible?	Yes / No	DS5	Approximately what proportion of the image is covered by above-ground vegetation such as shrubs and trees?	1-25% / 25-50% / 50-75% / 75-100% / None
PD6	Are there standing dead trees that could contribute to high-intensity burning?	Yes / No	DS6	What percentage of the RGB image is obstructed by smoke?	1-25% / 25-50% / 50-75% / 75-100% / No smoke
PD7	Are there isolated heat sources far from the main fire perimeter?	Yes / No / No fire	DS7	What percentage of the full scene exceeds 400 degrees Celsius?	<2% / 2-4% / 4-6% / >6% / None
PD8	How many emergency vehicles are visible in the scene?	0 / 1-2 / 3-4 / >4	DS8	What percentage of the full scene exceeds 200 degrees Celsius?	<5% / 5-10% / 10-15% / >15% / None
Classification			Cross-Modal Reasoning		
CL1	What is the dominant fire behavior observed in the scene?	Active fire / Smoldering / Extinguished / No fire	CMR1	What is the level of tree canopy obstruction of the fire's base?	Fully / Partially / Not obstructed / No fire
CL2	What is the dominant vegetation type in the scene?	Coniferous / Deciduous / Grassland / Shrubland	CMR2	What is the primary limitation to observing the active burn area in this scene?	Smoke / Canopy / Viewpoint / No major limitations / No fire
CL3	Which moisture level best describes the live vegetation?	Lush/Green / Transitioning / Dry/Cured	CMR3	What is the level of smoke obstruction of the fire's base?	Fully / Partially / Not obstructed / No fire
CL4	What is the density of the forest canopy?	Dense/Closed / Moderate / Sparse/Open / No forest	CMR4	What is the temperature of the hottest part of the fire in this scene in degrees Celsius?	100-200 / 200-300 / 300-400 / 400-500 / >500 / No hotspots
CL5	What is the primary fuel type on the ground by overall coverage?	Grass / Forest litter / Shrubs / Mixed	Flight Planning		
CL6	How accessible is the active fire area via roads or trails?	Clear / Partially / No clear access / No fire	FP1	What is the camera's viewing angle?	Nadir (top-down) / Oblique (angled)
Localization/Direction			FP2	What is the estimated flight altitude category?	0-50 m / 50-100 m / 100-150 m / >150 m
LD1	Where is the most intense hotspot located within the frame?	TL / TR / BL / BR / Center / No hotspots	FP3	What is the current level of safety risk of the UAV's position near flames or smoke?	High risk / Medium risk / Low risk / No fire
LD2	Where is the densest vegetation located?	TL / TR / BL / BR / Center / Uniform / No veg	FP4	At the UAV's current flight altitude, which scene feature is the biggest risk to safe or consistent UAV movement?	A. Rugged terrain B. Uneven forest C. Smoke columns D. No obstacles
LD3	From which region of the image does the largest smoke plume originate?	TL / TR / BL / BR / Center / Spread / No smoke			
LD4	What is the primary location of the man-made structures?	TL / TR / BL / BR / Center / No structures			

Prompt:

You are analyzing aerial wildfire imagery captured by a UAV. Two aligned images of the same scene are provided.

- The first image is a standard **RGB aerial image**.
- The second image is a **radiometric thermal image rendered with an inferno colormap**.

Use **both images together** to understand wildfire activity in the scene.

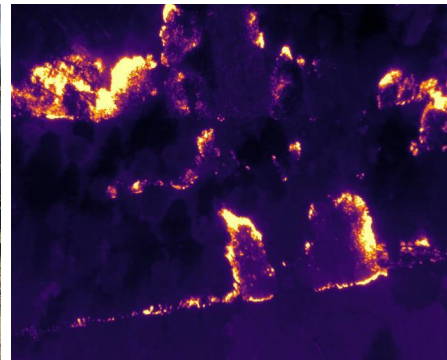
Temperature Summary (°C):

- Minimum Temp: 32.1
- Maximum Temp: 612.5
- Mean Temp: 96.4
- Top 3% Mean: 428.2

Use these statistics as anchor points to relate thermal colors to approximate temperatures.



RGB Image



Thermal Image

Question:

What is the spatial distribution of active hotspots in the scene?

- A. No active hotspots
- B. Concentrated hotspots
- C. Scattered hotspots
- D. Linear fireline

Example Model Responses

Qwen3-VL-8B: The thermal image reveals several spatially separated regions of elevated temperature distributed across the scene. These hotspots do not form a continuous line and are not concentrated in a single cluster.

Answer: C. Scattered hotspots ✓

LLaVA-v1.6-7B: The RGB image shows flames concentrated in one region, and the thermal map appears dominated by a single hotspot.

Answer: B. Concentrated hotspots ✗

InternVL2-8B: The thermal pattern forms an elongated region of heat suggesting a linear fire structure.

Answer: D. Linear fireline ✗

MiniCPM-V2: The thermal visualization does not clearly indicate strong hotspots above background levels.

Answer: A. No active hotspots ✗

Figure 1. Example WildFireVQA prompt and model responses. The prompt contains aligned RGB and thermal images, a radiometric temperature summary derived from the thermal TIFF, and a multiple-choice question. Different multimodal language models may interpret the scene differently based on visual and thermal cues.