FLOODMARK: A BENCHMARK FOR FLOOD SCENE UNDERSTANDING AND REASONING IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) are increasingly used to interpret complex visual scenes and support domain-specific decision making. To comprehensively evaluate their capabilities in flood-scene imagery understanding, we introduce **FloodMark**, a benchmark for fine-grained, evidence-aware flood hazard assessment. FloodMark contains 1,400 real-world flood and post-flood images and over 16,000 human-annotated image question pairs across 15 expert-authored tasks, covering Evidence Quality, Scene Context, Flood-state Perception, Physical Reasoning, and Operational Reasoning. The benchmark evaluates whether models recognize when visual evidence is insufficient to support a reliable conclusion. We evaluate 13 recent proprietary and open-weight MLLMs. Gemini 3.8 Flash achieves the highest overall accuracy of 73.2%, showing substantial room for improvement. More importantly, when the image does not provide enough evidence for a reliable answer, on average, models still give a specific answer in 70.0% of cases. This tendency can lead to unsupported conclusions about flood conditions and their operational consequences. These findings show that current MLLMs remain unreliable for fine-grained, evidence-aware flood-scene assessment.

1 INTRODUCTION

Flooding poses an immediate challenge to urban mobility and emergency response. Roadway-related incidents have been linked to 40–60% of flood fatalities in several developed countries (Panakkal & Padgett, 2024; Han & Sharif, 2020; Ashley & Ashley, 2008). Reliable situational awareness therefore depends on timely information about which roads are flooded, how severe the flooding is, and whether affected roads remain passable.

Much of this information is already captured visually. Transportation agencies operate large traffic-camera networks; for example, Houston TranStar publishes feeds from more than 700 cameras (Panakkal & Padgett, 2024). However, manually inspecting large numbers of camera feeds during a flood is difficult to scale, while existing automated pipelines often reduce roadway conditions to coarse flood or passability labels (Panakkal & Padgett, 2024). Recent vision-language models (VLMs) have begun to use such imagery for flood and disaster assessment. BayFlood applies VLMs to large-scale dashboard-camera imagery for urban flood detection (Franchi et al., 2026), while DisasTeller uses large vision-language models for post-disaster assessment and response planning (Chen et al., 2026). Multimodal large language models (MLLMs) extend this setting, raising the question of whether they can support more detailed flood-scene assessment.

Reliable flood assessment requires more than recognizing that flooding is present. It requires identifying visible hazards, estimating physical flood states such as water depth and flow, reasoning about response-relevant consequences such as road passability and damage, and recognizing when the image does not provide enough evidence for a reliable conclusion. Fine-grained multimodal evaluation has similarly been developed in medicine, finance, sports, and autonomous driving (Hu et al., 2024; Gan et al., 2025; Xia et al., 2025; Sima et al., 2025). Related work on reliable visual question answering further shows that answerability is itself an important capability: real-world questions may be unanswerable from the available image, and models must sometimes withhold an answer rather than return an unsupported one (Gurari et al., 2018; Whitehead et al., 2022; Miyai et al., 2025).

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.

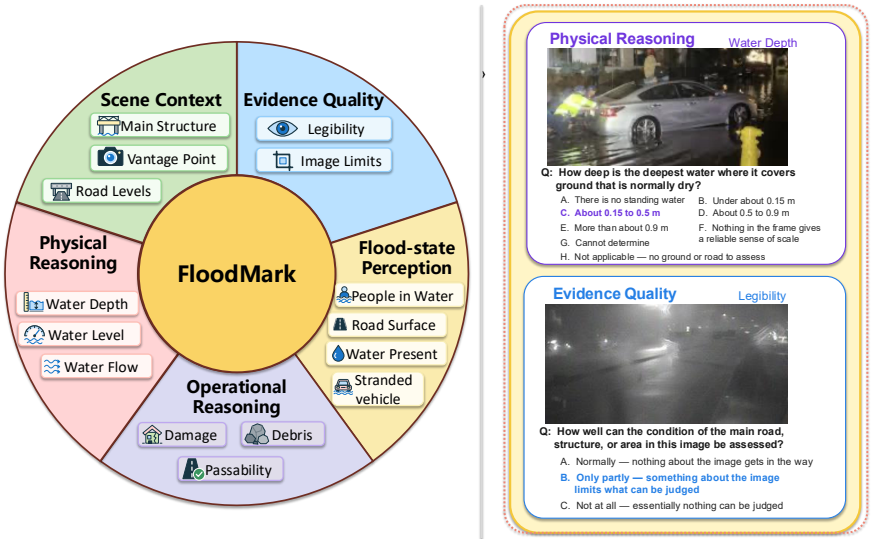


Figure 1: Overview of FloodMark, showing the five assessment dimensions, the 15 flood-assessment tasks, and representative image–question examples.

Existing disaster and flood benchmarks do not fully capture these requirements. Many widely used benchmarks rely primarily on Unmanned Aerial Vehicle (UAV) or satellite imagery (Rahnemoonfar et al., 2021; Sarkar & Rahnemoonfar, 2023; Wang et al., 2025a; Zhang et al., 2026), while ground-level and fixed-camera views remain underrepresented. Other flood focused benchmarks evaluate visual question answering, damage assessment, or reasoning segmentation (Sun et al., 2023; Bhattacharjya et al., 2026). However, existing benchmarks do not jointly evaluate physically grounded flood states such as depth, relative water level, and flow together with response-relevant judgments such as road passability, debris obstruction, and damage. They also do not explicitly distinguish insufficient visual evidence from a supported negative observation.

To address these gaps, we introduce **FloodMark**, a benchmark for fine-grained, evidence-aware flood hazard assessment. FloodMark organizes flood-scene assessment into five dimensions—Evidence Quality, Scene Context, Flood-state Perception, Physical Reasoning, and Operational Reasoning—and instantiates them as 15 expert-authored tasks. The benchmark contains 1,400 real-world flood and post-flood images and over 16,000 human-annotated image–question pairs. The tasks span the progression from establishing reliable visual evidence, to recognizing observable flood conditions, to reasoning about their physical state and operational consequences, as illustrated in Figure 1. FloodMark also treats insufficient visual evidence as a valid outcome through explicit *Cannot determine* answers.

We evaluate 13 recent proprietary and open-weight MLLMs on FloodMark. The best model, Gemini 3.8 Flash (DeepMind, 2026), achieves 73.2% overall accuracy. Averaged across models, performance is substantially lower on Physical Reasoning (45.7%) and Evidence Quality (48.4%) than on Scene Context (74.5%) and Flood-state Perception (70.5%). Across 13 models, they also fail to abstain in 70.0% of cases where the benchmark indicates that the image does not support a determinate answer. These results reveal substantial limitations in fine-grained and evidence-aware flood assessment. Our contributions are threefold:

- We propose a structured framework for fine-grained flood-scene assessment, covering Evidence Quality, Scene Context, Flood state Perception, Physical Reasoning, and Operational Reasoning, with explicit treatment of insufficient visual evidence.
- We introduce **FloodMark**, including 1,400 real-world flood and post-flood images and over 16,000 human-annotated image question pairs across 15 expert-authored flood tasks.
- We benchmark 13 recent proprietary and open-weight MLLMs and provide diagnostic analyses of task difficulty, real-world capture conditions, abstention, and model errors.

2 Related Work

2.1 Visual Understanding and Situational Awareness for Disaster and Flood Assessment

Visual and multimodal methods have been widely used to extract situational information from social media, satellite, aerial, and ground-level imagery during disasters. Early work focuses on humanitarian information extraction, damage classification, and large-scale damage mapping by combining visual, textual, and remote-sensing signals (Alam et al., 2018; Agarwal et al., 2020; Hao & Wang, 2020; Gupta et al., 2019). These approaches support tasks such as identifying affected infrastructure, estimating damage severity, and mapping disaster impacts for emergency response.

Flood-specific work further targets roadway conditions and other operationally relevant scene properties. Prior studies estimate floodwater level from surveillance or crowdsourced imagery, detect passable roads from remote-sensing and social-media data, and model how flood depth affects road disruption (Moy de Vitry et al., 2019; Chaudhary et al., 2019; Ahmad et al., 2019; Pregnolato et al., 2017). More recent systems combine traffic cameras, physical sensors, social data, and predictive models for real-time roadway situational awareness (Panakkal & Padgett, 2024). This line of work highlights both the value and the difficulty of extracting reliable road conditions from real-world observations: traffic-camera imagery can provide timely evidence, but image quality and source reliability vary, and manual inspection is difficult to scale. Recent vision-language approaches extend these capabilities to street-level flood detection, flood-depth estimation, and multimodal post-disaster assessment (Franchi et al., 2026; Liu et al., 2026; Chen et al., 2026; Xiao et al., 2026).

2.2 Reliable Visual Question Answering and Abstention

Prior work has studied when visual question answering systems should refrain from giving an answer. VizWiz has naturally occurring unanswerable questions caused by real-world image and question quality (Gurari et al., 2018). Reliable VQA studies selective answering, where models abstain when they are unlikely to answer correctly (Whitehead et al., 2022). Recent work such as MM-UPD evaluates whether multimodal models can detect unsolvable multiple-choice problems, including missing answers, incompatible choices, and mismatched image–question pairs (Miyai et al., 2025).

FloodMark differs in how abstention is defined and grounded. Rather than treating answerability as generic model uncertainty or an artificially unsolvable input, FloodMark includes insufficient evidence directly in expert-authored flood-assessment answer spaces. For absence-based tasks, it further distinguishes a supported negative observation from insufficient visual evidence. This allows abstention to be evaluated as part of flood-scene assessment itself.

2.3 Multimodal Benchmarks for Disaster and Flood Assessment

Flood-focused visual question answering has been explored through several datasets with different assessment targets. FloodNet, RescueNet-VQA, and FFD-IQA evaluate questions involving flooded conditions, roads, buildings, damage, risk, and spatial understanding from UAV or flood imagery (Rahnemoonfar et al., 2021; Sarkar & Rahnemoonfar, 2023; Sun et al., 2023). These benchmarks establish flood and post-disaster VQA as an evaluation setting, but primarily focus on visible scene conditions and damage-related questions.

Recent benchmarks broaden disaster evaluation toward multi-hazard and multi-stage reasoning. DisasterVQA, DisasterM3, and DisasterBench cover multiple disaster types and evaluate combinations of situational awareness, damage assessment, spatial reasoning, causal analysis, and response-oriented reasoning (Al-Mohannadi et al., 2026; Wang et al., 2025a; Zhang et al., 2026). FloodReasonBench further studies flood-specific language-conditioned reasoning segmentation rather than question answering (Bhattacharjya et al., 2026).

However, existing benchmarks leave several gaps for fine-grained flood assessment. Most rely primarily on UAV or satellite imagery, while ground-level and traffic-camera views that expose roadway depth and passability are less represented. They also do not jointly evaluate physically anchored flood states such as depth, relative water level, and flow together with downstream operational consequences. Finally, existing flood and disaster VQA benchmarks do not explicitly distinguish



Figure 2: FloodMark benchmark construction pipeline. Four steps cover real-world image collection, expert task and answer-option design, independent human annotation, and quality control with adjudication. The task space follows a three-stage, five-dimension assessment framework.

supported negative observations from cases where the visual evidence is insufficient. FloodMark addresses these gaps through a fixed expert-authored assessment space spanning flood-state perception, physical reasoning, operational reasoning, and evidence sufficiency. A detailed comparison is provided in Appendix A.

3 Benchmark Construction

FloodMark is constructed through four main steps: real-world image collection, expert task and answer-option design, independent human annotation, and quality control with adjudication. Within the task-design stage, we formulate flood-scene understanding as a structured assessment problem spanning three stages, five complementary dimensions, and 15 expert-authored tasks. The resulting benchmark contains 1,400 real-world flood and post-flood images and over 16,000 human-annotated image–question pairs. Figure 2 summarizes the overall construction pipeline.

3.1 Assessment Framework

Flood-scene assessment can be viewed as a progression from observing the hazard to understanding its impacts, consistent with prior work on flood-related situational awareness (Panakkal & Padgett, 2024). We extend this view by explicitly considering whether the image itself provides sufficient evidence for a reliable judgment.

We organize flood-scene assessment into three stages and five complementary dimensions, summarized in Table 1. The grouping reflects each task's role in the assessment process rather than a strict distinction between directly observed and inferred quantities.

The first stage, *Visual Basis*, establishes whether the image can support reliable assessment and what scene context is being observed. It contains *Evidence Quality*, which measures the reliability and limitations of the available visual evidence, and *Scene Context*, which captures the structure, viewpoint, and roadway configuration of the scene.

The second stage, *Observable State*, captures the visible flood situation and direct exposure. Its *Flood-state Perception* dimension includes water presence and roadway coverage, as well as stranded vehicles and people in flooded areas.

The final stage, *Inference*, characterizes the physical properties and response-relevant consequences of the observed scene. *Physical Reasoning* covers water depth, relative water level, and flow, while *Operational Reasoning* covers passability, debris obstruction, and damage. Although some of these tasks rely on visible scene evidence, they require mapping that evidence to response-relevant judg-

Table 1: FloodMark task taxonomy. The 15 tasks instantiate three stages and five complementary dimensions of flood-scene assessment.

Stage	Dimension	Task	Assessment Target & Interpretation
Visual Basis	Evidence Quality	Legibility	Reliability of scene assessment
		Image Limits	Visual factors that limit interpretation
	Scene Context	Main Structure	Primary structure or surface being assessed
		Vantage Point	Observation viewpoint
		Road Levels	Roadway configuration and multiple levels
Observable State	Flood-state Perception	Water Present	Water on normally dry ground
		Road Surface	Roadway flood condition
		Stranded Vehicles	Vehicles disabled or abandoned in floodwater
		People in Water	People present in flooded areas
Inference	Physical Reasoning	Water Depth	Floodwater depth inferred from visual cues
		Water Level	Water level relative to a reference object
		Water Flow	Still, flowing, or fast/turbulent water
	Operational Reasoning	Passability	Whether and what type of vehicle can pass
		Debris	Presence and severity of roadway obstruction
		Damage	Visible road, bridge, utility, and structural damage

ments rather than simply recognizing objects or conditions. For example, Passability and Debris require assessing whether a vehicle can traverse the observed roadway condition.

3.2 Task Construction

Starting from the five assessment dimensions, we operationalize flood-scene assessment as 15 targeted multiple-choice tasks. The task space is designed by researchers with expertise in civil, flood hazards, and disaster assessment. For each task, we specify the assessment target, decision rule, question, and complete answer space before annotation. The annotation rules are documented in a shared guide, and the full question text and answer options are provided in Appendix B. The resulting benchmark contains 13 single-choice tasks and 2 multi-select tasks.

The answer spaces are designed to represent meaningful flood conditions and response states. For ordinal quantities, the options follow the natural structure of the target. For example, Water Depth uses ordered depth intervals, while Passability distinguishes roads that are accessible to any vehicle, only to high-clearance vehicles, or to no vehicle. Road Surface separates dry, wet, partially flooded, and fully covered roadways, while Damage distinguishes roadway, bridge, utility, cosmetic building, and structural building damage.

All task definitions, questions, and answer options are authored by the domain experts. We do not use language models to generate questions or distractors. This avoids automatically generated language shortcuts and exploitable answer-choice artifacts reported in prior work (Burgess et al., 2025; Balepur et al., 2024).

Evidence sufficiency is incorporated directly into the answer options. When an image does not provide enough information for a reliable judgment, the model may answer "Cannot determine." For absence-based tasks, we further separate supported absence from insufficient evidence. For example, "No one visible—and the view is good enough that I would see them" represents a supported negative observation, whereas "Cannot determine" indicates that the visual evidence is insufficient. This distinction allows FloodMark to evaluate whether the image supports the model's conclusion, in addition to whether the conclusion itself is correct.

3.3 Image Collection

We collect 1,400 flood and post-flood images from multiple sources and viewpoints, including images from public websites, fixed traffic cameras, and aerial images. We retain a wide range of real-world capture conditions rather than restricting the dataset to visually clean images. The dataset includes low-resolution and nighttime images, as well as images affected by glare, lens obstruction, and occlusion. Such conditions are common in real-world flood monitoring and allow us to evaluate whether models can make reliable predictions when visual evidence is limited. For traffic-camera imagery, we remove the timestamp and camera-name bar before evaluation. This prevents models from using location or time metadata that is directly visible in the source image.

3.4 Human Annotation and Quality Control

FloodMark is annotated by five annotators from the authors, including four PhD students and one undergraduate student. Each image is independently labeled by two annotators using Label Studio.[1] The annotation interface is provided in Appendix D. Annotations with matching labels are directly retained as the gold answer. Disagreements are reviewed by a third annotator, who provides an independent label. When two of the three annotations agree, the majority label is used as the gold answer. Questions without a resolved gold label are not included in the benchmark. We assess annotation consistency using Gwet's AC1 (Gwet, 2008). Across the 15 tasks, the average agreement is AC1$= 0.80$. As a result, the benchmark contains over 16,000 human-annotated image-question pairs.

4 Experiments

We evaluate 13 recent MLLMs, including both proprietary and open-weight models, under a unified zero-shot evaluation protocol on FloodMark to assess overall flood-scene understanding and performance across the five assessment dimensions introduced in Section 3.

4.1 Models

We evaluate 13 MLLMs, including both proprietary and open-weight models. The proprietary models are Gemini 3.8 Flash (DeepMind, 2026), Claude Opus 5 (Anthropic, 2026), and GPT-5.6 Sol (OpenAI, 2026). The open-weight models include Kimi K3 (Team et al., 2026), GLM-5.3 Flash (GLM-5-Team et al., 2026), DeepSeek V4.1 Flash (DeepSeek-AI et al., 2026), Qwen3.5-397B-A17B (Qwen Team, 2026), STEP3-VL-10B (Huang et al., 2026), Phi-4-Reasoning-Vision-15B (Aneja et al., 2026), InternVL3.5-38B and InternVL3.5-8B (Wang et al., 2025b), Mistral-Small-3.1-24B[2], and Keye-VL-1.5-8B (Yang et al., 2025).

We report one primary setting for each model. When a model supports a configurable reasoning mode, we use its default reasoning setting in the main evaluation. Open-weight models are served locally, while the remaining models are accessed through their corresponding APIs.

4.2 Evaluation Setup

All models are evaluated in a zero-shot setting on the full FloodMark benchmark. Each instance contains one image and one multiple-choice question. We use the same instruction for all models and ask them to answer using only the visual evidence in the image.

For single-choice tasks, a model output answer is scored by exact match. For the two multi-select tasks, a model output answer is correct only when the predicted option set exactly matches the gold set. We use the same answer-extraction rules across models and score only the final answer.

We use accuracy as the primary evaluation metric and report both overall performance and results across the five assessment dimensions defined in Section 3.

5 Results

5.1 Benchmark Results

Table 2 summarizes the performance of the 13 evaluated models on FloodMark. Gemini 3.8 Flash achieves the highest accuracy at 73.2%, followed by Kimi K3 at 68.9% and Claude Opus 5 at 68.6%. GLM-5.3 Flash and DeepSeek V4.1 Flash reach 67.1% and 65.1%, respectively. No evaluated model exceeds 74% accuracy, indicating substantial room for improvement on fine-grained flood-scene understanding. Open-weight models are also competitive: Kimi K3 closely matches Claude Opus 5, while GLM-5.3 Flash and DeepSeek V4.1 Flash outperform GPT-5.6 Sol.

[1] https://labelstud.io/

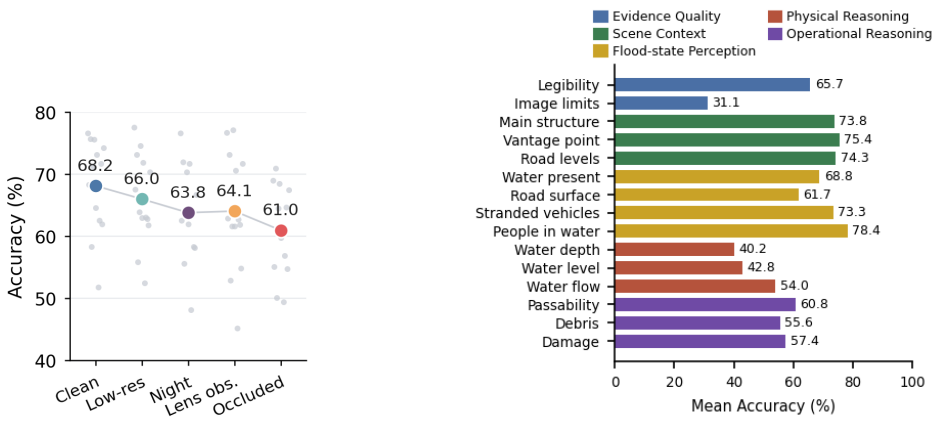
[2] https://mistral.ai/news/mistral-small-3-1/

Table 2: Performance on FloodMark across the five assessment dimensions. Higher is better.

Model	Evidence Quality	Scene Context	Flood-state Perception	Physical Reasoning	Operational Reasoning	Accuracy
Proprietary Models						
Gemini 3.8 Flash	**57.7**	**82.7**	**81.7**	**68.2**	67.8	**73.2**
Claude Opus 5	47.4	81.9	78.3	55.7	**69.3**	68.6
GPT-5.6 Sol	43.0	78.2	67.3	46.5	60.2	60.7
Open-weight Models						
Kimi K3	**57.1**	**81.3**	**75.6**	**58.0**	**66.1**	**68.9**
GLM-5.3 Flash	55.9	77.8	75.0	56.6	63.8	67.1
DeepSeek V4.1 Flash	49.8	78.2	73.9	49.7	65.7	65.1
Qwen3.5-397B-A17B	47.3	74.4	68.3	41.7	62.5	60.3
STEP3-VL-10B	53.1	66.4	70.3	40.4	57.0	58.6
InternVL3.5-38B	52.2	68.4	67.7	44.3	53.2	58.2
Phi-4-Reasoning-Vision-15B	41.3	72.1	70.7	44.0	52.6	58.1
Mistral-Small-3.1-24B	37.5	69.9	65.2	37.7	50.3	54.0
Keye-VL-1.5-8B	44.2	70.1	67.3	20.6	46.5	51.3
InternVL3.5-8B	43.1	67.0	55.8	30.4	38.0	47.7

Performance varies substantially across the five assessment dimensions. Averaged across all models, Scene Context reaches 74.5% accuracy and Flood-state Perception reaches 70.5%. Accuracy decreases to 57.9% on Operational Reasoning, 48.4% on Evidence Quality, and 45.7% on Physical Reasoning. This pattern shows that current MLLMs are considerably stronger at understanding scene context and directly visible flood conditions than at reasoning about flood physics or judging the reliability of the available visual evidence.

The task-level results show the same trend. As shown in Figure 3b, Image Limits is the most difficult task, with an average accuracy of 31.1%, followed by Water Depth at 40.2% and Water Level at 42.8%. In contrast, People in Water reaches 78.4%, while Vantage Point and Road Levels reach 75.4% and 74.3%, respectively. Overall, the largest performance gaps arise in evidence interpretation and fine-grained physical flood assessment. Full per-model results across all 15 tasks are provided in Appendix C.



(a) Performance across capture conditions.

(b) Performance across the 15 FloodMark tasks.

Figure 3: Diagnostic analysis on FloodMark. **Left:** accuracy across naturally occurring capture conditions on determinate downstream cases. Gold *Cannot determine* examples and the two Evidence Quality tasks are excluded to isolate the effect of capture conditions on downstream flood-scene understanding. Gray points denote individual models and colored points denote the mean across all 13 models. **Right:** mean accuracy across all 13 models on each of the 15 FloodMark tasks, grouped by assessment dimension.

5.2 Evidence Quality and Abstention

Real-world flood imagery often contains poor visibility, low resolution, occlusion, and other capture limitations. A direct comparison across capture conditions can conflate two different sources of difficulty: answering a question when sufficient visual evidence is available, and recognizing when the image does not support a reliable answer. We therefore first exclude examples whose gold answer is *Cannot determine*, and analyze abstention separately below.

For this analysis, we further exclude Legibility and Image Limits because these two Evidence Quality tasks directly assess image usability and degradation. Including them would mix the ability to recognize degraded imagery with the effect of that degradation on downstream flood-scene understanding and reasoning. We therefore evaluate the remaining 13 tasks to examine model performance when the scene supports a determinate answer.

As shown in Figure 3a, average accuracy across the 13 models is 68.2% on clean images, compared with 66.0% on low-resolution images, 63.8% at night, 64.1% under lens obstruction, and 61.0% on occluded images. These results show that degraded visual conditions remain a substantial challenge even when the scene contains sufficient evidence for a determinate answer, with the largest performance gap observed under occlusion.

We next consider the complementary case in which the visual evidence is insufficient to support a determinate answer. Across all 13 models, only 30.0% of gold *Cannot determine* cases are correctly identified; in the remaining 70.0%, models fail to abstain and select a non-abstaining option. By contrast, over-abstention on determinate cases is much lower, averaging 3.8%. This indicates that the dominant failure is not excessive abstention, but failing to recognize when the available visual evidence is insufficient.

As shown in Table 3, the three proprietary models exhibit higher correct-abstention rates than most evaluated open-weight models. Gemini 3.8 Flash, Claude Opus 5, and GPT-5.6 Sol reach 58.7%, 47.5%, and 49.0%, respectively. Across the three proprietary models, correct abstention averages 51.7%, compared with 23.4% across the ten open-weight models. This reveals a substantial difference in abstention behavior between the evaluated proprietary and open-weight models.

Together, these results show two complementary limitations of current MLLMs: performance degrades when visual evidence becomes harder to interpret, and models often fail to abstain when that evidence becomes insufficient.

Table 3: Abstention Analysis. Correct abstention: among gold *Cannot determine* cases, the model predicts *Cannot determine*. False certainty: among gold *Cannot determine* cases, the model predicts another option. Over-abstention: among other gold cases, the model predicts *Cannot determine*.

	Proprietary			*Open-weight*										*Mean*		
Metric	**Gemini**	**Opus**	**GPT-5.6**	**Kimi**	**GLM**	**DeepSeek**	**Qwen**	**STEP3**	**Phi-4**	**InternVL-38B**	**Mistral**	**Keye**	**InternVL-8B**	**Prop.**	**Open**	**All**
Correct Abst. ↑	58.7	47.5	49.0	31.3	39.2	30.5	24.8	20.2	10.5	18.6	25.9	14.7	18.5	51.7	23.4	30.0
False Cert. ↓	41.3	52.5	51.0	68.7	60.8	69.5	75.2	79.8	89.5	81.4	74.1	85.3	81.5	48.3	76.6	70.0
Over-Abst. ↓	6.8	4.6	8.1	3.7	2.9	3.9	0.9	4.6	0.6	0.5	2.5	4.0	5.7	6.5	2.9	3.8

5.3 Error Analysis

To better understand the failure modes of current MLLMs, we conduct a manual analysis of 300 incorrect predictions from six models: Gemini 3.8 Flash, Claude Opus 5, GPT-5.6 Sol, Kimi K3, GLM-5.3 Flash, and DeepSeek V4.1 Flash. We include all three proprietary models and the three strongest open-weight models on FloodMark. For each model, we randomly sample 50 incorrect predictions, stratified evenly across the five assessment dimensions with 10 examples per dimension. Annotators are shown the image, question, gold answer, and model prediction, with model identity hidden, and assign each case to one primary error category. Appendix E shows error examples and case studies.

Error taxonomy. We define six observable error categories. Our taxonomy combines general visual-grounding failures, such as unsupported claims and perception errors, with flood-specific failures in physical-state reasoning and operational inference. **Evidence Assessment Error** occurs when the model incorrectly judges whether the image provides enough visual evidence to answer

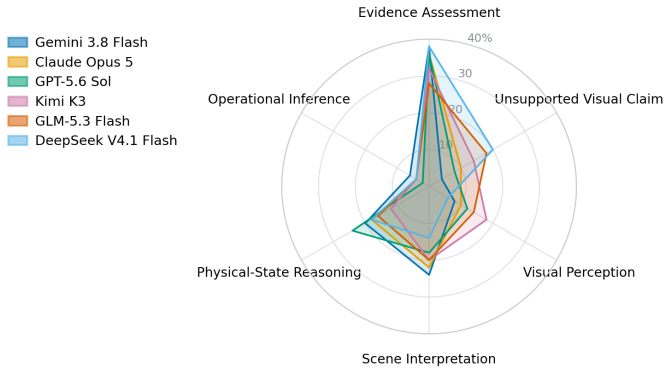


Figure 4: Error distribution of six representative MLLMs.

the question. **Unsupported Visual Claim** occurs when the model predicts an object, condition, or visible event that is not supported by the image. **Visual Perception Error** occurs when the model misses or misidentifies visual evidence that is clearly present. **Scene Interpretation Error** captures incorrect interpretation of the overall scene, viewpoint, roadway configuration, or other spatial structure. **Physical-State Reasoning Error** occurs when sufficient visual evidence is available but the model incorrectly assesses water depth, relative water level, or flow. Finally, **Operational Inference Error** occurs when sufficient scene evidence is available but the model incorrectly assesses passability, debris obstruction, or damage.

Figure 4 shows the error patterns of the six analyzed models. Evidence Assessment Error is the largest primary error category for every model, accounting for 28–38% of the analyzed errors. This shows current MLLMs consistently struggle with evidence assessment in real-world flood scenes. Physical-State Reasoning Error is also substantially more common than Operational Inference Error. Across all 300 cases, the two categories account for 18.0% and 4.0%, respectively. The former involves fine-grained judgments of water depth, relative water level, and flow, while the latter concerns broader scene-level outcomes such as passability, obstruction, and damage.

The remaining error distributions vary across models. Unsupported Visual Claims account for a larger share of the analyzed errors for the three open-weight models (17.3%) than for the three proprietary models (7.3%). Overall, evidence assessment is a consistent weakness across all six models, while scene interpretation, physical-state reasoning, visual perception, and unsupported visual claims contribute differently across individual models.

6 Conclusion

We introduce **FloodMark**, a benchmark for fine-grained and evidence-aware flood-scene understanding in MLLMs. Across evaluated models, performance is substantially stronger on scene context and directly visible flood conditions than on physical reasoning and evidence assessment, while models also frequently fail to abstain when visual evidence is insufficient. Together, these findings reveal a gap between recognizing what is visible in a flood scene and producing a physically grounded conclusion that is supported by the available evidence. Closing this gap is important for developing multimodal models that can be used reliably in real-world flood and disaster assessment.

AI Use Statement

In this work, we used generative AI tools for several tasks requiring disclosure: assisting with the implementation of methods, specifically for code debugging and troubleshooting; assisting with translation; and supporting discussions of experimental results and their potential interpretations. All interpretations and conclusions reported in the paper were independently evaluated and determined by the authors based on the experimental evidence. We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, or support qualitative or thematic data anal-

ysis. The remaining required-disclosure tasks, including those related to theoretical or conceptual model development, mathematical claims and proofs, hypothesis formulation, and research methodology or experimental design, were not applicable to this work. Additionally, we used generative AI tools to assist with drafting and editing portions of the manuscript for clarity and readability. All AI-assisted code, translations, discussions, and manuscript text were reviewed by the authors. In particular, AI-assisted debugging was validated by executing and checking the resulting code. We take responsibility for the final content of this work, including text, claims, code, results, and other artifacts produced with the aid of generative AI.

REFERENCES

Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 346–353, 2020.

Kashif Ahmad, Konstantin Pogorelov, Michael Riegler, Olga Ostroukhova, Pål Halvorsen, Nicola Conci, and Rozenn Dahyot. Automatic detection of passable roads after floods in remote sensed and social media data. *Signal Processing: Image Communication*, 74:110–118, 2019. ISSN 0923-5965. doi: https://doi.org/10.1016/j.image.2019.02.002. URL https://www.sciencedirect.com/science/article/pii/S0923596518311536.

Aisha Al-Mohannadi, Ayisha Firoz, Yin Yang, Muhammad Imran, and Ferda Ofli. Disastervqa: A visual question answering benchmark dataset for disaster scenes, 2026. URL https://arxiv.org/abs/2601.13839.

Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters, 2018. URL https://arxiv.org/abs/1805.00713.

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. Phi-4-reasoning-vision-15b technical report, 2026. URL https://arxiv.org/abs/2603.03975.

Anthropic. System card: Claude opus 5. Technical report, 2026. URL https://www-cdn.anthropic.com/ceaf5c7ff2783855203fde8208ec311252dced5b/Claude%20Opus%205%20System%20Card.pdf.

Sharon T. Ashley and Walker S. Ashley. Flood fatalities in the united states. *Journal of Applied Meteorology and Climatology*, 47(3):805 – 818, 2008. doi: 10.1175/2007JAMC1611.1. URL https://journals.ametsoc.org/view/journals/apme/47/3/2007jamc1611.1.xml.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10308–10330, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.555. URL https://aclanthology.org/2024.acl-long.555/.

Rajat Bhattacharjya, Yoomee Jung, Minwoo Kim, Sing-Yao Wu, Eli Bozorgzadeh, Nalini Venkatasubramanian, and Nikil Dutt. Floodreasonbench: Benchmarking vlm reasoning segmentation for embodied flood response at the edge, 2026. URL https://arxiv.org/abs/2608.15410.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research, 2025. URL https://arxiv.org/abs/2503.13399.

P. Chaudhary, S. D'Aronco, M. Moy de Vitry, J. P. Leitão, and J. D. Wegner. Flood-water level estimation from social media images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W5:5–12, 2019. doi: 10.5194/isprs-annals-IV-2-W5-5-2019. URL https://isprs-annals.copernicus.org/articles/IV-2-W5/5/2019/.

Zhaohui Chen, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, and Daniel Dias-da Costa. Integration of large vision language models for efficient post-disaster damage assessment and reporting. *Nature Communications*, 17(1):1481, 2026. doi: 10.1038/s41467-025-68216-z. URL https://doi.org/10.1038/s41467-025-68216-z.

Google DeepMind. Gemini 3.8 flash - model card — google deepmind, 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

DeepSeek-AI, :, Anyi Xu, B. Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, C. C. Yu, Chao Jin, Chaofan Lin, Chen Dong, Chenbing Wang, Chenfan Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyuan Zhang, Chenhao Xu, Chenqi Zhao, Chenze Shao, Chuhao Wang, Chuqi Zhang, Damai Dai, Dejian Yang, Deli Chen, Di Huang, Di Wu, Donghao Li, Erhang Li, Eric Fu, F. Zhou, Fangwei Zhou, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanglin Li, Guanting Chen, Guoai Cao, Guofan Fan, Guolai Meng, Guowei Li, Haichuan Zhang, Haiyang Ma, Haiyang Shen, Han Li, Han Yu, Han Zhang, Hangyuan Deng, Hanwei Xu, Hanxiang Xu, Hanxun Zhong, Hao Guo, Hao Jiang, Hao Li, Hao Qin, Haodong Wen, Haofen Liang, Haofeng Huang, Haohua Liu, Haoling Zhang, Haoming Luo, Haoran Yang, Haotian Xu, Haotian Yuan, Haoting Huang, Haowen Luo, Haoyang Cai, Haoyu Chen, Haozhe Ji, Hengran Zhang, Hengrui Wang, Hengxu Wu, Honghui Ding, Hongxuan Tang, Huadong Wang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J. Yang, J. H. Jin, J. H. Zhang, J. X. Zou, Jia Yu, Jiahui Zhou, Jiajun Chen, Jialiang Huang, Jialin Zhao, Jiamin Tang, Jian Zhou, Jianan Tong, Jianwen Li, Jiaqi Zhu, Jiarui Wang, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiaying Ding, Jibai Lu, Jiewen Hu, Jin Yan, Jincheng Zhai, Jingchang Chen, Jingcheng Hu, Jingli Zhou, Jingsheng Xu, Jingting Xiang, Jingyan Yun, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jinpeng Wang, Jinyi Chen, Jinyi Hu, Jiping Yu, Jueliang Guo, Junbo Pei, Junbo Sun, Junguang Jiang, Junjie Qiu, Junkang Zhou, Junqi Liu, Junren Li, Junxian Li, Junxiao Song, Junyi Guo, Kai Dong, Kaifeng Chen, Kaige Gao, Kang Guan, Kangdong Yuan, Ke Hong, Ke Xu, Kefan Zhao, Kexin Ji, Kexin Zhang, Kexing Zhou, Kuai Yu, Lan Zhang, Lean Wang, Lecong Zhang, Lei Wang, Letian Gao, Liang Zhao, Liansheng Xu, Lihua Guo, Lingxiao Luo, Lingyue Fu, Litao Deng, Litong Wang, Liyue Zhang, Longhao Chen, Lu Chen, Luotian Huang, Luyao Ma, Luyao Wang, M. S. Di, Max Mei, Menghao Ye, Miao Cui, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingjing Zhang, Mingqi Wei, Mingshu Chen, Mingxing Liu, Mingxu Zhou, Mingyu Xu, Mingyu Yang, Mingze Wang, Muyang Chen, Ni Shentu, Ning Wang, Niufang Ning, Panpan Huang, Peixin Cong, Peiyi Wang, Peiyuan Xin, Pengfei Ren, Pengfei Yan, Pengle Zhang, Qi Kang, Qi Tang, Qiancheng Wang, Qiang Li, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qizhou Guo, Rongxian Xu, Rui Ding, Rui Hu, Rui Tian, Rui Yu, Ruidong Zhu, Ruifan Xu, Ruihan Yang, Ruihang Xia, Ruijie Lu, Ruilin Geng, Ruipeng Hong, Ruiqi Ge, Ruisong Zhang, Ruize Sun, Ruizhe Pan, Runji Wang, Runqian Chen, Runxin Xu, Ruohong Tian, Ruomeng Shen, Ruoyu Zhang, Ryan X., S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoyuan Chen, Shengding Hu, Shengkai Lin, Shengwen Ran, Shengyu Liu, Shengyuan Jia, Shi Bai, Shi Feng, Shicheng Xu, Shichun Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shiyuan Feng, Shufan Gong, Shuhan Lin, Shuiping Yu, Shunfeng Zhou, Shuo Yang, Shuomeng Wang, Shuting Guo, Shuting Pan, Shuying Yu, Sinuo Cao, Siyi Lin, Sizhe Chen, Songyang Chen, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tongrui Xiong, Wangding Zeng, Wei Liu, Wei Zhang, Weibin Xu, Weihao Zeng, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Shao, Wenkai Yang, Wenli Zhang, Wenlu Wang, Wenlve Huang, Wenqian Yan, Wentao Zhang, Xi Gao, Xiang He, Xiang Li, Xiangli Li, Xiangwen Wang, Xiangying Zhang, Xiankui Wei, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojian Qu, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaoyao Zou, Xiaoyuan Li, Xicheng Guo, Xieting Chu, Xin Cheng, Xin Liu, Xin Xie, Xinbo Xu, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xintong Yao, Xinyang Chen, Xinyong Jiang, Xinyu Yang, Xinyu Yang, Xu Chen, Xuanyu Wang, Xubei Zhong, Xuecheng Su, Xuejie Liu, Xuheng Lin, Xujie Fan, Xuncheng Zhao, Xuwei Fu, Y. C. Yan, Y. H. Jiang, Y. T. Wu, Y. W. M., Y. Z. Wang, Yafei Gao, Yang Yang, Yang Zhang, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Meng, Yao Zhao, Yaofeng Sun, Yaohui Wang,

Yaoyang Ye, Yehang Yin, Yexinrui Wu, Yi Qian, Yi Tao, Yi Yu, Yichao Zhang, Yichen Jiang, Yicheng Wang, Yifan Ding, Yifan Shi, Yifeng Peng, Yifeng Zhai, Yijia Wu, Yiliang Xiong, Yilun Wang, Ying He, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yiping Wang, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyao Yang, Yiyuan Liu, Yizai Cai, Yizhen Wei, Yizhi Wang, Yonglun Yang, Yongqi Zhuo, Yongqiang Guo, Yongtong Wu, Yu Wu, Yu Zhang, Yuan Bian, Yuan Cheng, Yuan Ou, Yuan Sun, Yuanfan Xu, Yuanhang Sun, Yuanhao Li, Yuchen Liu, Yuchen Yao, Yudong Han, Yuduan Wang, Yuhan Wu, Yuhao Meng, Yuheng Zou, YuKun Li, Yunchuan Wang, Yunfan Xiao, Yunfan Xiong, Yupeng Chen, Yuqian Cao, Yuqian Wang, Yuqing Chen, Yushun Zhang, Yutong Lin, Yuwei Xiao, Yuxian Gu, Yuxiang Chen, Yuxiang Huang, Yuxiang Luo, Yuxiang You, Yuxin Chen, Yuxin Xiang, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhe Guo, Yuzhen Huang, Yuzhuo Bai, Z. Y. Z., Zanlin Ni, Zehao Wang, Zehua Zhao, Zehui Ren, Zejun Zhao, Zhangli Sha, Zhanying Wang, Zhaochen Zhang, Zhaoshuai Du, Zhe Fu, Zhean Xu, Zhenda Xie, Zheng Liu, Zhengyan Zhang, Zhenhua Dong, Zhewen Hao, Zhibang Wang, Zhibin Gou, Zhicheng Ma, Zhihao Li, Zhihong Shao, Zhihuan Huang, Zhijie Li, Zhirui Lu, Zhixian Huang, Zhixuan Chen, Zhixuan Chen, Zhixuan Pan, Zhiyu Wu, Zhizhou Ren, Zhu He, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zili Zhang, Zilin Li, Zilong Hou, Zilong Lyu, Ziqiao Wang, Ziwei Xie, Ziya Zhang, Ziyi Gao, Zizheng Pan, Zonglin Li, Zongqing Yao, Zui Chen, Zuofan Wu, Chenchen Ling, Chengyu Hou, Chong Chen, D. Li, Di Qi, Dongjie Ji, Fang Wei, Fanyi Xia, Fei Xie, Feiyi Tan, Hailong Guo, Haiyan Zhai, Hui Zhou, Huihui Tan, Huijie Li, Jia Luo, Jia Song, Jialu Cai, Jian Liang, Jiangting Zhou, Jiaqi Gao, Jiayi Shao, Jie Chen, Jieyu Yang, Jin Chen, Jingde Zhang, Jingzi Zhou, Jinqian Wang, Jinyang Liu, JinZhao Sun, Junhua Ling, Junmin Zheng, Kaicheng Yang, Ke Xu, Le Su, Leyi Xia, Liangfeng Ding, Lin Zhuo, Linwang Ma, Linyan Zhu, Liyu Cai, Luqi Yao, M. K. Zhang, Meng Li, Miao Lin, Miaojun Wang, Min Zhang, Mingming Li, Mingming Wang, Mingze Yin, Minmin Han, Nan Cao, Ning Wang, Ningxin Ma, Panpan Wang, Peihan Lin, Peng Sun, Peng Zhang, Qian Ying, Qiang Xiang, Qiao Wang, Qingmiao Mao, Qiwei Jiang, Rongli Jin, Ruyi Chen, Sha Tao, Shangmian Sun, Shaoqing Wu, Shichao Zou, Si Lei, Tianyang Zhang, Tianyu Sun, Tingting Yin, W. L. Xiao, Wei An, Wei Li, Wei Wang, Weiwei Lin, Wenqing Hou, X. Lin, Xiangfei Meng, Xianzhu Huang, Xiao Peng, Xiaoqian Li, Xiaoting Zhang, Xiaowen Sun, Xiaoxiang Wang, Xiaoyu Ye, Xinrou Zhang, Xinyu Zhang, Xue Cao, Xueyin Chen, Yanan Zhou, Yanhong Xu, Yao Xia, Yao Xu, Yi Shao, Yihong Zhang, Yiling Ma, Ying Tang, Yining Lou, Yiru Chen, Yishi Piao, Yixuan Chen, Yong Xiong, Yuchen Xuan, Yuehan Yang, Yuer Xu, Yukun Zha, Yunxian Ma, Yuping Lin, Yuting Yan, Yutong Xie, Yuwen Sheng, Yuxuan Zhu, Zekai Zhang, Zhe Ju, Zhenzhen Lin, Zheren Gao, Zheyang Sun, Zhigang Yan, Zhongyu Wu, Zi Wang, Zihua Qu, Ziling Yan, and Ziyi Wan. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Matt Franchi, Nikhil Garg, Wendy Ju, and Emma Pierson. Improving flood detection with large-scale dashboard camera data. *Nature Communications*, 17(1):9208, 2026. doi: 10.1038/s41467-026-75938-1. URL https://doi.org/10.1038/s41467-026-75938-1.

Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12867–12874, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758230. URL https://doi.org/10.1145/3746027.3758230.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu,

Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery, 2019. URL https://arxiv.org/abs/1911.09296.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018. URL https://arxiv.org/abs/1802.08218.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: https://doi.org/10.1348/000711006X126600. URL https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/000711006X126600.

Zhongyu Han and Hatim O. Sharif. Vehicle-related flood fatalities in texas, 1959–2019. *Water*, 12 (10), 2020. ISSN 2073-4441. doi: 10.3390/w12102884. URL https://www.mdpi.com/2073-4441/12/10/2884.

Haiyan Hao and Yan Wang. Leveraging multimodal social media data for rapid disaster damage assessment. *International Journal of Disaster Risk Reduction*, 51:101760, 2020. ISSN 2212-4209. doi: https://doi.org/10.1016/j.ijdrr.2020.101760. URL https://www.sciencedirect.com/science/article/pii/S2212420920312620.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm, 2024. URL https://arxiv.org/abs/2402.09181.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, Jingcheng Hu, Kangheng Lin, Liang Zhao, Mitt Huang, Song Yuan, Wenwen Qu, Xiangfeng Wang, Yanlin Lai, Yingxiu Zhao, Yinmin Zhang, Yukang Shi, Yuyang Chen, Zejia Weng, Ziyang Meng, Ang Li, Aobo Kong, Bo Dong, Changyi Wan, David Wang, Di Qi, Dingming Li, En Yu, Guopeng Li, Haiquan Yin, Han Zhou, Hanshan Zhang, Haolong Yan, Hebin Zhou, Hongbo Peng, Jiaran Zhang, Jiashu Lv, Jiayi Fu, Jie Cheng, Jie Zhou, Jisheng Yin, Jingjing Xie, Jingwei Wu, Jun Zhang, Junfeng Liu, Kaijun Tan, Kaiwen Yan, Liangyu Chen, Lina Chen, Mingliang Li, Qian Zhao, Quan Sun, Shaoliang Pang, Shengjie Fan, Shijie Shang, Siyuan Zhang, Tianhao You, Wei Ji, Wuxun Xie, Xiaobo Yang, Xiaojie Hou, Xiaoran Jiao, Xiaoxiao Ren, Xiangwen Kong, Xin Huang, Xin Wu, Xing Chen, Xinran Wang, Xuelin Zhang, Yana Wei, Yang Li, Yanming Xu, Yeqing Shen, Yuang Peng, Yue Peng, Yu Zhou, Yusheng Li, Yuxiang Yang, Yuyang Zhang, Zhe Xie, Zhewei Huang, Zhenyi Lu, Zhimin Fan, Zihui Cheng, Daxin Jiang, Qi Han, Xiangyu Zhang, Yibo Zhu, and Zheng Ge. Step3-vl-10b technical report, 2026. URL https://arxiv.org/abs/2601.09668.

Zhangding Liu, Neda Mohammadi, and John E. Taylor. Knowledge-guided vision-language inference for image-based urban flood depth estimation, 2026. URL https://arxiv.org/abs/2509.04772.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Helen Li, Ziwei Liu, and Kiyoharu Aizawa. Unsolvable problem detection: Robust understanding evaluation for large multimodal models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6497–6540, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.326. URL https://aclanthology.org/2025.acl-long.326/.

M. Moy de Vitry, S. Kramer, J. D. Wegner, and J. P. Leitão. Scalable flood level trend monitoring with surveillance cameras using a deep convolutional neural network. *Hydrology and Earth System Sciences*, 23(11):4621–4634, 2019. doi: 10.5194/hess-23-4621-2019. URL https://hess.copernicus.org/articles/23/4621/2019/.

OpenAI. Openai gpt-5.6 system card. Technical report, 2026. URL https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf.

Pranavesh Panakkal and Jamie Ellen Padgett. More eyes on the road: Sensing flooded roads by fusing real-time observations from public data sources. *Reliability Engineering & System Safety*, 251: 110368, 2024. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2024.110368. URL https://www.sciencedirect.com/science/article/pii/S095183202400440X.

Maria Pregnolato, Alistair Ford, Sean M. Wilkinson, and Richard J. Dawson. The impact of flooding on road transport: A depth-disruption function. *Transportation Research Part D: Transport and Environment*, 55:67–81, 2017. ISSN 1361-9209. doi: https://doi.org/10.1016/j.trd.2017.06.020. URL https://www.sciencedirect.com/science/article/pii/S1361920916308367.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. doi: 10.1109/ACCESS.2021.3090981.

Argho Sarkar and Maryam Rahnemoonfar. Rescuenet-vqa: A large-scale visual question answering benchmark for damage assessment. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1150–1153, 2023. doi: 10.1109/IGARSS52108.2023.10281747.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 256–274, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72943-0.

Yimin Sun, Chao Wang, and Yan Peng. Unleashing the potential of large language model: Zero-shot vqa for flood disaster scenario, 2023. URL https://arxiv.org/abs/2312.01882.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun

Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, and Naoto Yokoya. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response, 2025a. URL https://arxiv.org/abs/2505.21089.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pp. 148–166, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20058-8. doi: 10.1007/978-3-031-20059-5_9. URL https://doi.org/10.1007/978-3-031-20059-5_9.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan fang Wang, Weining Shen, and Hanjie Chen. Sportu: A comprehensive sports understanding benchmark for multimodal large language models, 2025. URL https://arxiv.org/abs/2410.08474.

Yiming Xiao, Kai Yin, and Ali Mostafavi. Crisisense-rag: Crisis sensing multimodal retrieval-augmented generation for rapid disaster impact assessment. *Computer-Aided Civil and Infrastructure Engineering*, 48:100096, 2026. ISSN 1093-9687. doi: https://doi.org/10.1016/j.cacaie.2026.100096. URL https://www.sciencedirect.com/science/article/pii/S1093968726030823.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Guowang Zhang, Han Shen, Hao Peng, Haojie Ding, Hao Wang, Haonan Fan, Hengrui Ju, Jiaming Huang, Jiangxia Cao, Jiankang Chen, Jingyun Hua, Kaibing Chen, Kaiyu Jiang, Kaiyu Tang, Kun Gai, Muhao Wei, Qiang Wang, Ruitao Wang, Sen Na, Shengnan Zhang, Siyang Mao, Sui Huang, Tianke Zhang, Tingting Gao, Wei Chen, Wei Yuan, Xiangyu Wu, Xiao Hu, Xingyu Lu, Yi-Fan Zhang, Yiping Yang, Yulong Chen, Zeyi Lu, Zhenhua Wu, Zhixin Ling, Zhuoran Yang, Ziming Li, Di Xu, Haixuan Gao, Hang Li, Jing Wang, Lejian Ren, Qigen Hu, Qianqian Wang, Shiyao Wang, Xinchen Luo, Yan Li, Yuhang Hu, and Zixing Zhang. Kwai keye-vl 1.5 technical report, 2025. URL https://arxiv.org/abs/2509.01563.

Tan Zhang, Quanyou Li, Lu Zhang, Jun Liu, Xiaofeng Zhu, and Ping Hu. Disasterbench: A multimodal benchmark for uav-based disaster response in complex environments, 2026. URL https://arxiv.org/abs/2606.06217.

A Comparison with Existing Disaster Benchmarks

Table 4 compares representative vision-language benchmarks for flood and disaster assessment.

Human-authored Q/A space indicates that question templates or tasks and answer options are directly specified by humans rather than generated by automatic models. **Flood-state taxonomy** denotes a fixed flood-specific task set that distinguishes multiple observable flood conditions across scene elements. **Multi-factor flood physics** explicitly evaluates depth, relative water level, and flow as separate flood-state quantities. **Operational reasoning** covers response-relevant judgments such as passability, roadway obstruction, and damage. **Evidence quality** explicitly evaluates image interpretability or whether sufficient visual evidence is available for assessment. **Explicit abstention** treats insufficient visual evidence as a valid gold answer.

Table 4: Comparison of representative vision-language benchmarks for flood and disaster assessment. For multi-disaster benchmarks, flood-subset statistics precede full-dataset statistics in parentheses when available; *n/r* indicates not reported.

Benchmark	Scope	Images	QA / Inst.	Human-authored Q/A Space	Flood-state Taxonomy	Multi-factor Flood Physics	Operational Reasoning	Evidence Quality	Explicit Abstention
FloodNet (Rahnemoonfar et al., 2021)	Flood	2,343	∼4.5K	✓	✓	✗	✗	✗	✗
RescueNet-VQA (Sarkar & Rahnemoonfar, 2023)	Post-disaster	4,300	100K	✗	✗	✗	✓	✗	✗
FFD-IQA (Sun et al., 2023)	Flood	2,058	22,422	✗	✗	✗	✓	✗	✗
DisasterVQA (Al-Mohammadi et al., 2026)	Multi-disaster	289 (1,395)	851 (4,405)	✗	✗	✗	✓	✗	✗
DisasterM3 (Wang et al., 2025a)	Multi-hazard	n/r (26,988)	n/r (123,010)	✗	✗	✗	✓	✗	✗
DisasterBench (Zhang et al., 2026)	Multi-disaster	n/r (5,330)	n/r (29,300)	✗	✗	✗	✓	✗	✗
FloodMark	**Flood**	**1,400**	**16,403**	✓	✓	✓	✓	✓	✓

B Complete Questions and Answer Options

Table 5 lists the complete question stems and answer options used in FloodMark.

Table 5: Complete question stems and answer options for the 15 FloodMark tasks.

1. Main Structure What is the main built structure or surface in this image whose physical condition could be assessed?	A road or street A bridge A building A storage tank An open area with no prominent structure A wide view with several structures—no single one dominates Built structures are present, but destroyed or in ruins Nothing built is visible Cannot determine
2. Legibility How well can the condition of the main road, structure, or area in this image be assessed?	Normally—nothing about the image gets in the way Only partly—something about the image limits what can be judged Not at all—essentially nothing can be judged
3. Image Limits What about this image limits what can be read from it? Select every one that applies.	Nothing—the image is clear It is too dark Glare Rain, spray, dirt, or other obstructions on the lens Blur, compression artefacts, or low resolution Objects of interest are too small at this scale Something large is blocking the view The camera is pointed away from anything useful The scene has nothing to do with roads or buildings
4. Vantage Point From what vantage point was this image taken?	From the ground, at eye or vehicle level From a fixed pole or mast, looking down From the air, at an angle From directly overhead, looking straight down Some other vantage point Cannot determine
5. Road Levels Does this image show more than one roadway level, such as an elevated freeway or overpass above a ground-level road?	No—a single roadway level Yes—both visible, and their conditions differ Yes—both visible, conditions look the same Only an elevated roadway is visible No roadway is visible Cannot determine
6. Water Present Is there standing or flowing water on normally dry ground, on or beside the road, structure, or area being assessed?	Yes No Only water that belongs there is visible—ocean, bay, river, canal, pond or pool Cannot determine Not applicable—no road, structure or area to assess
7. Road Surface What is the state of the road surface being assessed? In a wide area view, answer for the most affected roads visible.	Dry Wet, but with no water standing on it Water standing on part of it—puddles, or one lane Water covering all the travel lanes in a continuous sheet Covered by sand, mud, or storm debris, with no standing water Not applicable—no road is visible Cannot determine

8. Water Depth How deep is the deepest water where it covers ground that is normally dry?	There is no standing water Under about 0.15 m About 0.15 to 0.5 m About 0.5 to 0.9 m More than about 0.9 m Nothing in the frame gives a reliable sense of scale Cannot determine Not applicable—no ground or road to assess
9. Water Level Where is the water relative to {object}, at its deepest point? {object} is a human-annotated specific reference object, which varies in different images.	Below it Level with it Above it It is completely under water
10. Water Flow Is the water on normally dry ground moving?	Still, or with wind ripples only Flowing Fast or turbulent Not applicable—no water on normally dry ground Cannot determine
11. Passability From what is visible alone, could a vehicle get through? In a wide area view, answer for the worst-affected stretch visible.	Yes, any vehicle Only a high-clearance vehicle, such as a pickup or SUV No Not applicable—no road is visible Cannot determine
12. Stranded Vehicles How many vehicles look disabled or abandoned in the water?	None visible—and the view is good enough that I would see them One Two to five More than five Cannot determine
13. People in Water Is anyone in standing or flowing water on normally dry ground?	No one visible—and the view is good enough that I would see them Yes, at least one person Cannot determine
14. Debris Is there debris or an obstruction on the road?	None visible—and the view is good enough that I would see it Minor debris—a vehicle could drive over it Something large enough to block a travel lane on its own Not applicable—no road is visible Cannot determine
15. Damage What damage is visible? Select every one that applies.	No damage visible—structures are clearly seen and appear intact Damage to the road surface The roadway has been washed out Damage to a bridge Downed power lines or poles Downed trees Exterior or cosmetic damage to a building only—siding, shingles, gutters, fences Structural damage to a building—a breached roof deck, a failed wall, partial or total collapse Some other damage Cannot determine

Table 6: Full task-level accuracy (%) of the 13 evaluated models on FloodMark. Tasks are grouped by the five assessment dimensions.

	Evidence Quality		Scene Context			Flood-state Perception				Physical Reasoning			Operational Reasoning			Overall
Model	Legi-bility	Image Limits	Main Structure	Vantage Point	Road Levels	Water Present	Road Surface	Stranded Vehicles	People in Water	Water Depth	Water Level	Water Flow	Passability	Debris	Damage	Acc.
Proprietary Models																
Gemini 3.8 Flash	61.6	53.7	78.8	85.3	84.1	84.3	80.7	86.9	74.7	70.8	54.4	79.5	81.5	52.8	69.1	73.2
Claude Opus 5	67.2	27.5	78.9	86.7	80.0	80.8	71.3	78.4	82.6	50.6	49.4	67.2	77.9	58.4	71.5	68.6
GPT-5.6 Sol	67.6	18.4	76.8	80.6	77.3	68.5	65.4	63.6	71.5	41.5	49.9	48.0	73.9	46.0	60.8	60.7
Open-weight Models																
Kimi K3	70.5	43.8	77.4	85.8	80.6	71.5	67.0	79.0	85.0	58.3	48.6	67.1	71.7	56.1	70.6	68.9
GLM-5.3 Flash	71.0	40.8	77.9	85.1	70.4	70.8	66.9	77.5	85.0	58.5	46.6	64.8	71.0	60.2	60.2	67.1
DeepSeek V4.1 Flash	69.2	30.3	72.9	85.0	76.7	69.1	69.1	76.9	80.6	45.0	48.9	55.2	71.7	54.4	71.1	65.1
Qwen3.5-397B-A17B	66.4	28.3	77.7	82.2	63.4	59.1	59.9	76.4	78.1	36.6	41.3	47.3	63.2	55.8	68.3	60.3
STEP3-VL-10B	64.2	41.9	60.3	63.4	75.4	69.7	60.2	74.2	77.1	38.8	31.2	51.1	56.1	58.3	56.8	58.6
InternVL3.5-38B	66.7	37.7	71.4	59.5	74.3	60.6	64.7	69.8	75.5	39.8	45.3	47.9	56.5	54.7	48.3	58.2
Phi-4-Reasoning-Vision-15B	58.1	24.4	76.5	62.5	77.3	69.0	51.8	79.5	82.3	32.7	43.8	55.4	51.8	69.1	36.8	58.1
Mistral-Small-3.1-24B	60.7	14.3	67.9	75.8	66.0	65.3	42.4	72.0	81.0	19.9	38.8	54.4	38.5	70.7	41.7	54.0
Keye-VL-1.5-8B	64.8	23.7	69.9	68.4	71.8	59.8	51.9	75.0	82.6	12.4	25.4	24.0	41.8	61.1	36.7	51.3
InternVL3.5-8B	66.5	19.6	72.5	60.1	68.5	66.0	50.4	44.2	62.8	17.8	33.2	40.1	35.1	24.5	54.5	47.7

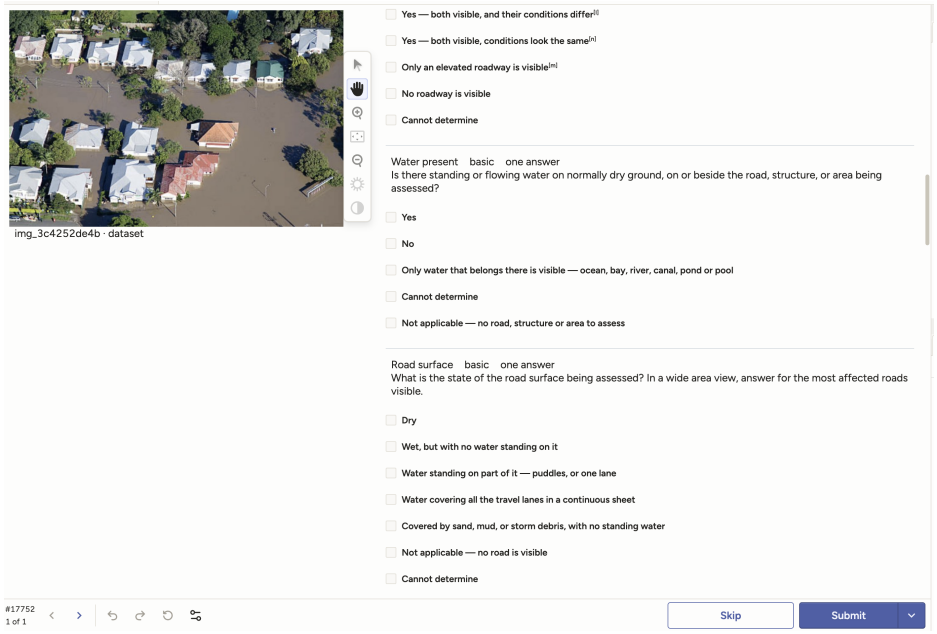


Figure 5: Annotation Interface

C Full Task-Level Results

Table 6 reports the complete task-level accuracy of all 13 evaluated models.

D Annotation Interface

Figure 5 shows the annotation interface for each annotator.

E Error Examples

Question: How many vehicles look disabled or abandoned in the water ?

Gold Answer : Cannot Determine

Model Answer: None visible - and the view is good enough that i would see them

Error Type : Evidence Assessment Error
Error Reason: The aerial view is too high and distant to make out individual vehicles, so whether any are in the water can't be determined. The model took "no vehicles seen" to mean "no vehicles present" and wrongly claimed the view was clear enough to support that.



Figure 6: Example of an Evidence Assessment Error.

Question: Is the water on normally dry ground moving?

Gold Answer : Not applicable – no water on normally dry ground

Model Answer: Flowing

Error Type: Visual Perception Error
Error Reason: The water in the image is flowing, but it's the river flowing in its own channel. There's no water on normally dry ground, so the answer should be "Not applicable." The model answered "Flowing" because it saw moving river water and didn't separate the river from land that has actually flooded. It answered whether the water moves, not whether water on dry ground moves.



Figure 7: Example of a Visual Perception Error.

Question: Does this image show more than one roadway level, such as an elevated freeway or overpass above a ground-level road?

Gold Answer: No — a single roadway level

Model Answer : Yes — both visible, and their conditions differ

Error Type: Scene Interpretation Error
Error Reason: The model took the pedestrian footbridge with railings to be an overpass, and the flooded channel below it to be a ground-level road. Both are wrong: the car is submerged in a stream beside a footbridge, and the image has no road on either level.



Figure 8: Example of a Scene Interpretation Error.

Question: How deep is the deepest water where it covers ground that is normally dry?

Gold Answer : There is no standing water

Model Answer: Under about 0.15 m

Error Type: Physical-State Reasoning Error
Error Reason: The street is wet and reflective from rain, but the water is only a thin film on the pavement. There's no standing water, and the curbs, tires and building bases are all dry above the surface. The model took the wet, shiny road as shallow floodwater and gave it a depth ("under about 0.15 m"). It read how reflective the surface was as water depth instead of checking whether any water had actually pooled.

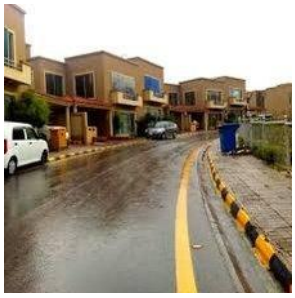


Figure 9: Example of a Physical-State Reasoning Error.

Question : Is there debris or an obstruction on the road?

Gold Answer : None visible — and the view is good enough that I would see it

Model Answer: Something large enough to block a travel lane on its own

Error Type : Operational Inference Error
Error Reason: The road is flooded, and the only large objects on it are two stalled cars. There's no debris. The model appears to have counted the stranded cars as a lane-blocking obstruction, which confuses vehicles caught in the flood with debris on the road. It judged debris severity wrongly even though the scene is visible enough to answer.



Figure 10: Example of an Operational Inference Error.