DisasterM3: A Remote Sensing Vision-Language Dataset for Disaster Damage Assessment and Response

Junjue Wang[1*], **Weihao Xuan**[1,2*], **Heli Qi**[2,3], **Zhihao Liu**[1], **Kunyi Liu**[3], **Yuhan Wu**[4], **Hongruixuan Chen**[1], **Jian Song**[2], **Junshi Xia**[2], **Zhuo Zheng**[5], **Naoto Yokoya**[1,2†]
[1]The University of Tokyo, [2]RIKEN AIP, [3]Waseda University,
[4]Stony Brook University, [5]Stanford University

Abstract

Large vision-language models (VLMs) have made great achievements in Earth vision. However, complex disaster scenes with diverse disaster types, geographic regions, and satellite sensors have posed new challenges for VLM applications. To fill this gap, we curate the first remote sensing vision-language dataset (DisasterM3) for global-scale disaster assessment and response. DisasterM3 includes 26,988 bi-temporal satellite images and 123k instruction pairs across 5 continents, with three characteristics: **1) Multi-hazard**: DisasterM3 involves 36 historical disaster events with significant impacts, which are categorized into 10 common natural and man-made disasters. **2) Multi-sensor**: Extreme weather during disasters often hinders optical sensor imaging, making it necessary to combine Synthetic Aperture Radar (SAR) imagery for post-disaster scenes. **3) Multi-task**: Based on real-world scenarios, DisasterM3 includes 9 disaster-related visual perception and reasoning tasks, harnessing the full potential of VLM's reasoning ability with progressing from disaster-bearing body recognition to structural damage assessment and object relational reasoning, culminating in the generation of long-form disaster reports. We extensively evaluated 14 generic and remote sensing VLMs on our benchmark, revealing that state-of-the-art models struggle with the disaster tasks, largely due to the lack of a disaster-specific corpus, cross-sensor gap, and damage object counting insensitivity. Focusing on these issues, we fine-tune four VLMs using our dataset and achieve stable improvements (up to 10.4%↑QA, 2.1↑Report, 40.8%↑Referring Seg.) with robust cross-sensor and cross-disaster generalization capabilities. Project: `https://github.com/Junjue-Wang/DisasterM3`.

1 Introduction

Onset natural and man-made disasters represent one of humanity's greatest challenges, causing devastating impacts across national borders [47, 8]. These catastrophic events (including earthquakes, tsunamis, floods, explosions, storms, etc) claim tens of thousands of lives globally each year while causing massive infrastructure damage and economic losses [33, 25]. Remote sensing (RS), as an ultra-long-distance Earth observation technology, has been widely used in disaster scenarios, i.e., hurricane damage assessment [29], landslide detection [36], mapping of burn area and ecological impacts [26], etc. Considering the urgency and timeliness of disaster relief, developing AI-based algorithms is necessary.

The recent advent of large vision-language models (VLMs) [16, 42, 6] has achieved substantial milestones in computer vision due to their exceptional ability to reason about visual and linguistic

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

clues and summarize high-level human-readable text. Inspired by the success of the generic domain, remote sensing has also explored the applications of VLMs, i.e., image classification [14], image captioning [13], visual question answering [41], etc. These remote sensing-tailored VLMs show great potential as general-purpose task solvers for multi-task scenarios. Unlike existing research that primarily addresses general geospatial tasks, our work explores the reasoning capabilities of VLMs in extreme disaster scenarios, thereby supporting rescue teams and planning personnel in making informed decisions.

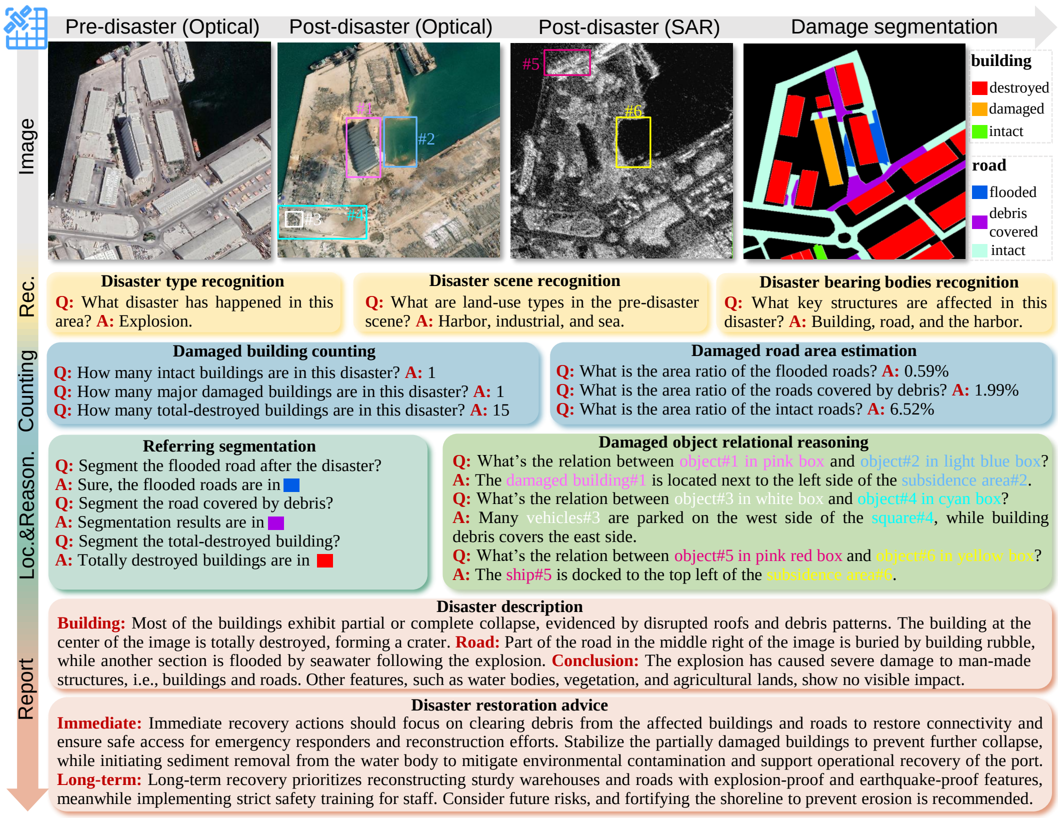

Figure 1: Task taxonomy in DisasterM3 dataset. Each scene includes the paired pre- and post-disaster images. The modalities of post-disaster images are optical `or` SAR. The 9 tasks derive from 5 essential capabilities for bi-temporal disaster assessment and response: recognition, counting, localization, reasoning, and report generation.

To facilitate the development of VLMs in disaster response, we propose the DisasterM3 dataset, featuring multi-hazard, multi-sensor, and multi-task challenges. As shown in Fig. 1, the DisasterM3 dataset includes the co-registered pre- and post-disaster optical and SAR images, as well as disaster instruction pairs. Our motivation is that a well-performing VLM should possess the ability to achieve a comprehensive understanding of disaster scenarios by responding to the instructions of rescuers. Based on this assumption, we build our task taxonomy by summarizing five essential capabilities required for disaster assessment and response: disaster recognition, damage counting, reasoning and localization, and disaster report generation. Then, these capabilities are delineated with 9 disaster-related tasks, carefully aligning with assessment and response requirements. The diversity of scenarios is ensured by meticulously collecting images from 36 disaster events covering 5 continents. A comprehensive data cleaning, annotation repurposing, instruction design, manual verification, and sampling pipeline is leveraged to generate high-quality annotations. After extensive benchmarking experiments, we found that the cutting-edge VLMs struggle on disaster tasks. To address this, four VLMs were fine-tuned using our dataset and achieve stable improvements across different tasks and sensors, providing solid baselines. The main contributions could be summarized as follows:

1. To advance intelligent disaster response, we introduce DisasterM3, the first multi-hazard, multi-sensor, and multi-task remote sensing dataset for vision-language understanding. It

includes 26,988 bi-temporal optical and SAR images, 123,010 instruction pairs across disaster-bearing body recognition, structural damage assessment, referring segmentation, object-relational reasoning, disaster captioning, and restoration advice generation.

2. To systematically analyze the efficacy of existing models on disaster tasks, we benchmark 14 advanced large VLMs, including open-source, commercial, and remote sensing methods. The comparative and detailed analysis illuminates their capabilities while identifying several critical directions for future improvement in disaster-focused vision-language understanding.
3. To provide strong baselines, we fine-tune Qwen2.5-VL, InternVL3, LISA, and PSALM on the DisasterM3 dataset, achieving consistent performance enhancements across all evaluation tasks and sensor modalities. With the injection of disaster corpus, the fine-tuned models exhibit good stability to prompt variations, serving as solid baseline solutions.

2 Related Work

General Vision-language Model. Assisted by the strong reasoning abilities of large language models, VLMs have transformed the visual perception domain by enabling the interpretation and reasoning about images through natural language interfaces. Several leading VLMs, including Flamingo [1], MiniGPT-4 [53], LLaVA [17], LLaVA-OneVision [16], InstructBLIP [6], and Qwen2-VL [42], have achieved remarkable results on vision-language tasks. However, these models are limited to generating only textual outputs that describe the image holistically. This restricts their applicability in damage assessment tasks that require the pixel-level detailed understanding. Several approaches have emerged to extend VLMs with fine-grained visual understanding. Ferret [49], Kosmos-2 [27], and VisionLLM [43] incorporate grounding functionalities through bounding box coordinate regression. Besides, LISA [15], PixelLM [31], GLaMM [30], and PerceptionGPT [28], integrate mask decoders to generate object masks from specialized tokens. For richer representation, PSALM [52] and HyperSeg [44] leverage queries in Mask2Former for unified segmentation. Despite their capabilities, generic VLMs exhibit substantial limitations in disaster scenarios due to insufficient domain-specific knowledge, restricting their operational utility in emergency response applications.

Table 1: Comparison of DisasterM3 with existing remote sensing vision-language datasets.

Dataset	Propose	#Optical	#SAR	#MT pairs*	#Text	Recognition	Counting	Localization	Reasoning	Caption
RSICD [23]	General	10,921	-	-	54,605	✓	✓	-	✗	✗
RSICap [11]	General	2,585	-	-	2,585	✓	✓	-	✗	✗
DIOR-RSVG [51]	General	17,402	-	-	38,320	✗	✗	Box	✗	✗
RRSIS-D [20]	General	17,402	-	-	17,402	✗	✗	Pixel	✗	✗
RSVQA-HR [22]	General	10,659	-	-	1,066,316	✓	✓	-	✗	✗
EarthVQA [41]	General	6,000	-	-	208,593	✓	✓	-	✗	✗
RSIEval [11]	General	100	-	-	933	✓	✓	-	✗	✗
VRSBench [18]	General	29,614	-	-	205,317	✓	✓	Box	✗	✗
XLRSBench [39]	General	1,400	-	-	45,942	✓	✓	Box	✓	✗
GeoChatSet [14]	General	106,747	-	-	308,861	✓	✓	Box	✓	✗
TeoChatlas [13]	General	351,957	-	245,210	554,071	✓	✓	Box	✗	✓
FloodNet [29]	Disaster	2,348	-	-	7,345	✓	✓	-	✗	✗
DisasterM3 (Ours)	Disaster	**22,214**	**4,774**	**15,881**	**123,010**	✓	✓	**Pixel**	✓	✓

* MT pairs (multi-temporal pairs) denote the number of pre/post-disaster image pairs.

Remote Sensing Vision-language Dataset. Following the substantial progress of general VLMs, the RS field has likewise undergone accelerated development, accompanied by the emergence of numerous specialized vision-language datasets [48, 46]. Focusing on holistic analysis, EarthVQA [41] and RSIEval [11] datasets provide manual instructions for visual question answering (VQA) and image captioning tasks. Leveraging GPT-4, VRSBench [18] introduced visual grounding tasks to evaluate the object reasoning abilities and XLRSBench [39] focuses on ultra-high-resolution image understanding. GeoChatSet [14] and TeoChatlas [13] collect the existing classification and detection datasets for secondary development, formulating the unified instruction-following datasets. Although TeoChatlas involves some disaster scenes, the instructions focus on common recognition tasks. FloodNet [29] is a VQA disaster dataset that assesses the buildings and roads affected by Hurricane Harvey. Limited by its single disaster and simple tasks, it is difficult to fully unleash the potential of VLMs. Overall, RS visual-language datasets for general geospatial tasks have reached a considerable level of maturity, yet there persists a notable deficiency in datasets addressing specialized geoscience challenges. For this case, we design the DisasterM3 dataset that is tailored for global disaster assessment and response with multi-sensor images, bi-temporal inputs, refined damage masks, and diverse visual understanding tasks in the context of disaster.

3 DisasterM3 Dataset

As shown in Fig. 2, we collect 36 historical natural and man-made significant disasters to construct the DisasterM3 dataset. There are 26 events from the xBD [9] and BRIGHT [5] dataset, we extend 10 new events using Maxar's Open Data program [24]. Considering these optical sensors (WorldView series) have similar spatial resolutions, all pre- and post-disaster images were pre-processed into 0.8 m. We collect the post-disaster Synthetic Aperture Radar (SAR) images from Capella Space [4] and Umbra [37]. Considering the amplitude data in the VV or HH bands, SAR images were terrain-corrected, stretched into [0, 255], and finally resampled to match the optical resolution. We performed the georeferencing to ensure that the pre- and post-disaster image pairs are strictly aligned spatially. Following the United Nations Satellite Centre (UNOSAT) Emergency Mapping Products [38], and the Federal Emergency Management Agency (FEMA) [7], we design 9 essential tasks required for disaster assessment and response, evaluating the VLM performances from different aspects.

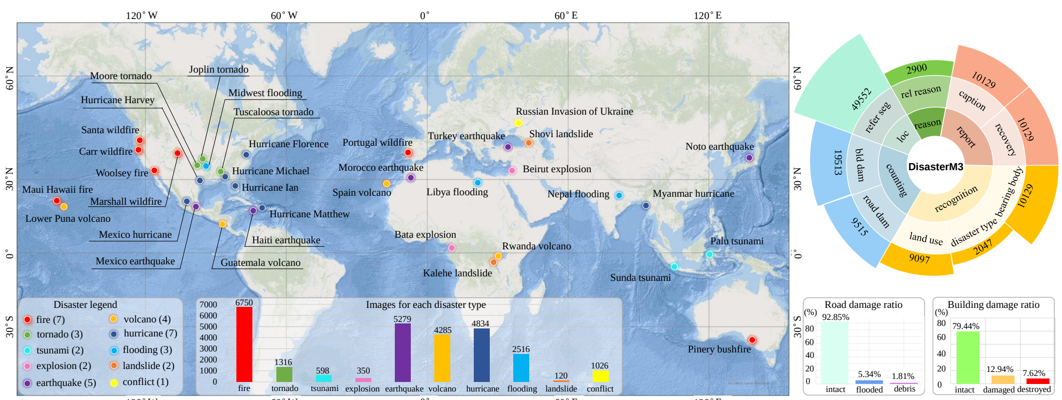

Figure 2: The DisasterM3 dataset involves 36 significant natural and man-made disaster events (10 types) across five continents. Diverse disaster-centric tasks provide a comprehensive evaluation benchmark for VLMs.

3.1 Perception and Reasoning Tasks in the Context of Disaster

Disaster Recognition. The disaster recognition tasks provide a brief description of disaster scenes, i.e., disaster types, land-use types and key disaster-bearing body. The disaster type follows the official definition and we chose 13 common land-use types from the AID dataset [45] for annotation. The land-use answers include: airport, bridge, river, forest, low vegetation, pond, parking, port, viaduct, residential area, industrial area, commercial area, and sea. Disaster-bearing bodies are the key resources that are damaged by disasters [8], and we focus on 12 types, i.e., building, stadium, open-space ground, bridge, dam, road, port facility, storage tank, farmland, forest, coastline, and mining area. Based on basic recognition types, users could have a rough disaster profile.

Damage Assessment. The damage assessment provides a quantitative analysis of disaster-bearing body. We chose the road and building, two important man-made structures for damage assessment. We annotate instance-level building damage masks using 'intact', 'damaged', and 'destroyed' types following FEMA guidelines. As a critical transportation hub, road accessibility plays a vital role in emergency response and recovery efforts. We classify the damaged roads into three types, i.e., 'intact', 'flooded (blocked by water)', and 'debris covered (blocked by debris)'. Based on these damage masks, the building counting and road area estimation instructions were automatically generated. The imbalanced sample distributions of damaged buildings and roads (Fig. 2) reveal the actual challenges for model optimization.

Disaster Referring Segmentation. Each disaster includes different forming factors and prone environments. In addition to disaster-bearing-body mapping, we identify the key visual objects and perform risk analysis using referring segmentation. As shown in Fig. 3, the first example shows an earthquake scene. In addition to referring segmentation for disaster-bearing body, we also design the task for finding the optimal rescue places shown in Fig 3(d). Similarly, Fig 3(e) shows the place where rescuers could find the available vehicles for dispatch. As for the volcano eruption scene,

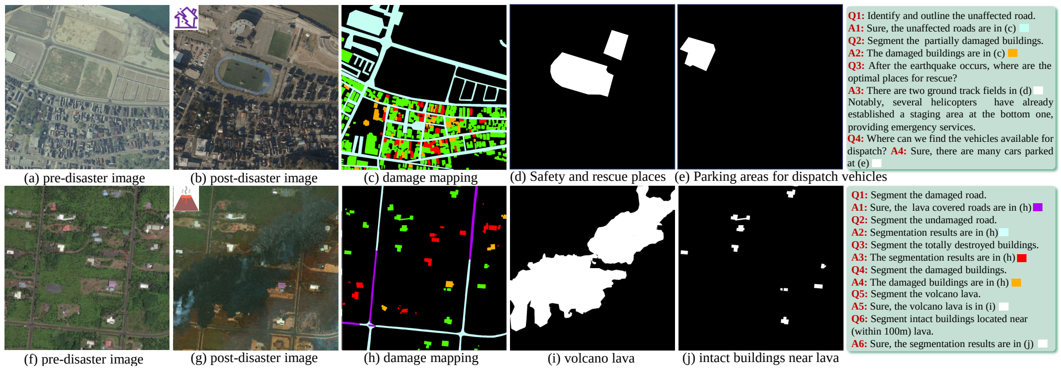

Figure 3: Disaster referring segmentation task involves disaster-bearing body mapping and risk analysis. By querying, rescuers could accurately locate the disaster-related objects.

we set the instruction tasks to individually map damaged buildings and roads, as well as the lava. Considering the situation, the intact buildings near the lava are also required for segmentation. By polygon distance analysis using the ArcGIS toolbox, the intact buildings within a 100-meter proximity to lava are segmented, providing early warning information. All the referring segmentation tasks are designed according to the specific disaster scenarios, which enable the rescuers to accurately locate the disaster-related objects and places.

Damaged Object Relational Reasoning. To capture the spatial relationships between multiple damaged objects, relational reasoning tasks are designed. In Fig. 4 wildfire scene, the spatial relationships between unaffected buildings and refuge squares, as well as between burnt grassland and unaffected trees, reveal crucial patterns in disaster response and spread prevention. The war conflict scene depicts the damaged industrial area, where the relationships between key facilities, factories, and transportation hubs are clarified. The reasoning task provides spatial analysis services for multiple objects, helping rescuers to understand critical facility spatial dependencies.

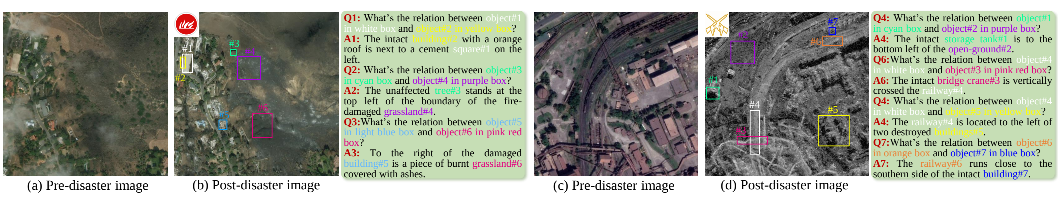

Figure 4: Damaged object relational reasoning task describes spatial relationships between key facilities, revealing crucial patterns in the object dependencies.

Disaster Comprehensive Report. To go beyond traditional perception tasks, the comprehensive reports are designed for the holistic analysis of disaster situations. Fig. 5 shows two samples of disaster caption and restoration advice. The earthquake caption describes the collapsed buildings and blocked roads, causing severe traffic congestion. Immediate response advice prioritizes the deployment of temporary shelters within the stadium for displaced survivors, a recommendation visibly implemented in the post-disaster image. Long-term recovery focuses on earthquake-resistant strategies in rebuilding and disaster protocols to mitigate seismic risks. The flooding caption summarizes that roads, buildings, and natural areas experienced severe inundation, while water bodies expanded and merged with flooded regions. Correspondingly, repairing critical transportation infrastructure, establishing temporary residential facilities, and implementing disease prevention protocols are proposed as immediate response measures. The installation of drainage systems integrated with local hydrological networks is recommended as a long-term strategy. Fig. 5 (e) and (f) shows the word cloud of disaster reports. Thanks to the wide range of disaster types, the words are diverse in terms of both nouns and verbs. Most words are disaster-centric, describing bearing bodies, damage impacts, response strategies, etc. Comprehensive disaster reports equip rescuers with enhanced situational awareness and evidence-based decision support.

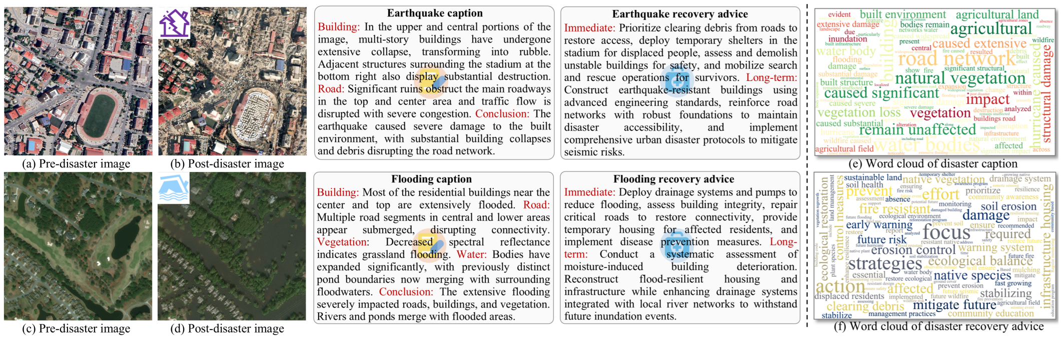

Figure 5: (Left) The disaster comprehensive reports provide a holistic analysis of disaster situations and evidence-based rescue support. It is notable that immediate earthquake response prioritizes deploying temporary shelters within the stadium for displaced survivors, an intervention demonstrated in the post-disaster image. (Right) Word cloud of reports shows that the disaster-centric words have a considerable degree of diversity.

3.2 Dataset Construction Pipeline

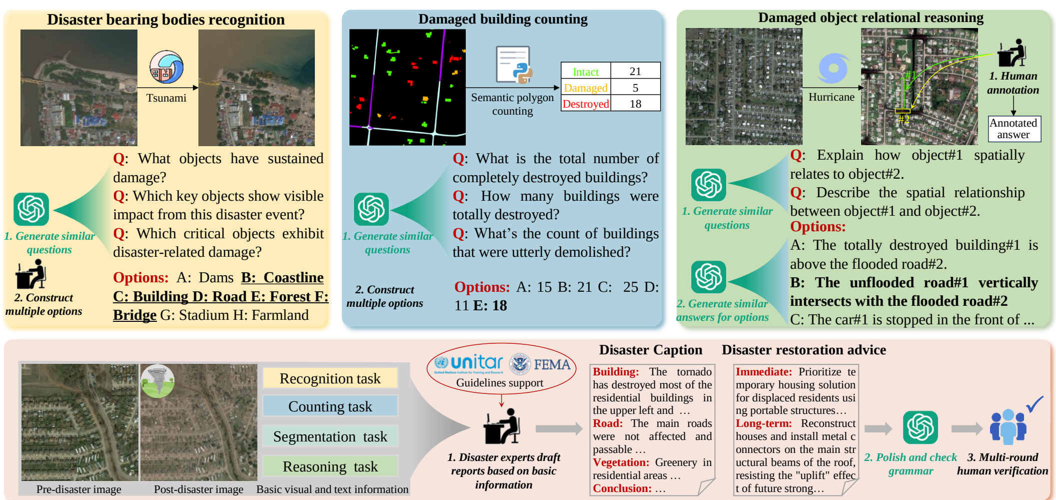

Figure 6: Dataset construction pipeline. We conduct a thorough process of question designing, answer annotation, and generation of similar questions as well as other options. Multi-round inspection controls the quality of each construction step.

Following the common vision-language data pipeline [21, 50], we divided the whole dataset into `Instruct` (17,190 Optical images, 3,798 SAR images, and 92,968 instruction pairs) and `Bench` sets (5,024 Optical images, 976 SAR images, and 30,042 instruction pairs). We describe the detailed annotation process in Fig. 6. As for recognition tasks, GPT-4o was employed to generate diverse prompt variations with similar semantic intent. Disaster domain experts subsequently annotate correct answers for these prompts. These question-answer pairs constitute the `Instruct` training set. To formulate the multiple-choice `Bench` set, correct answers were combined with other options. Regarding counting tasks, we counted semantic polygons using annotated road and building damage masks, generating correct answers. The similar options are generated with controlled deviations ($\pm 20\%$ and $\pm 40\%$) to maintain plausibility. As for relational reasoning, the experts annotate bounding boxes and describe the concrete relationship. We use GPT-4o to analyze the image by listing other significant relationships, generating alternatives. As for disaster reports, by referring to bi-temporal images and all basic task information, multiple experts draft the disaster caption and restoration advice following goals of UNITAR and FEMA projects. GPT-4o then polished the reports and corrected grammar errors. Finally, the multi-round verification was performed for controlling quality (Appendix §A). As for uninterpretable parts of SAR imagery, we annotate answers using co-registered `optical` images and then apply the instructions to SAR images. Using this pipeline, more future disasters can be effectively extended for DisasterM3 dataset.

4 Benchmark Experiments

Implementation Setting. As the DisasterM3 dataset features multi-sensor and multi-task, we comprehensively benchmark VLMs under four settings: Optical-Optical and Optical-SAR QA tasks, as well as Optical-Optical and Optical-SAR referring segmentation tasks.. We evaluate LLaVA-OneVision [16], InternVL3 [55], Kimi-VL [35], and Qwen2.5-VL [3] on QA tasks. In addition, we also tested commercial models such as GPT-4o [12] and Claude-3 [2] for comparison with the open-source models. As for remote sensing VLMs GeoChat [14], TeoChat [13], EarthDial [34] are chosen for evaluation. As for referring segmentation models, generic VLMs models such as LISA [15], PSALM [52], and HyperSeg [44], alongside the remote sensing model GeoPixel [32] were benchmarked. We fine-tuned Qwen2.5-VL-7B, InternVL3-8B, LISA and PSALM on our `Instruct` set. Model details are provided in Appendix §B.

Evaluation Metrics. Following common settings [16, 34], we adopted accuracy (%) for the multiple-choice tasks, i.e., disaster scene recognition (DSR), disaster type recognition (DTR), bearing body recognition (BBR), damaged building counting (DBC), damaged road estimation (DRE), object relational reasoning (ORR). The open-ended tasks are scored using GPT-4.1 at a scale of 5 points. Disaster caption is measured from damage assessment precision (DAP), damage detail recall (DDR), and factual correctness (FC). Restoration advice is measured from recovery necessity (RN), strategic completeness (SC), and action priority precision (APP). The average accuracy (AVG) denotes the overall performance. Evaluation prompts are provided in Appendix §C. As for referring segmentation, we chose cIoU and mIoU following previous work [15, 52].

4.1 Comparative Results

Domain gap for disaster scenarios. Tab. 2 presents performance evaluations on optical-optical settings for QA tasks. As a traditional VLM, LLaVA-1.5 exhibited significant limitations when processing disaster scenes due to the domain gap. By leveraging extensive multi-modal pretraining datasets and implementing the AnyRes architecture, LLaVA-OV demonstrates enhancements in both accuracy and multi-image processing capabilities. As efficient Mixture-of-Experts (MoE) VLMs, Kimi-VL-A3B-Think exceeds Kimi-VL-A3B-Instruct in mathematical counting tasks (BDC, DRE). However, the non-negligible domain gap limits their application on complex tasks, particularly degrading performance to near-random levels on the ORR task. This motivated our development of the DisasterM3 dataset, which identifies performance gaps through the `Bench` set while providing complementary training data via the `Instruct` set.

Table 2: Benchmarking results of VLMs on DisasterM3 `Bench` set with optical-optical setting.

Method	Accuracy (%)							Disaster Caption				Restoration Advice			
	AVG	DSR	DTR	BBR	BDC	DRE	ORR	AVG	DAP	DDR	FC	AVG	RN	APP	SC
Random Guess	-	-	*20*	-	*20*	*20*	*20*	-	-	-	-	-	-	-	-
• Open-source models															
LLaVA-1.5-7B [19]	12.1	4.2	-	-	-	-	20.0	-	-	-	-	-	-	-	-
LLaVA-OV-7B [17]	24.5	16.3	53.5	3.7	26.4	24.2	22.7	1.66	1.50	1.53	1.93	2.30	3.01	2.08	1.81
Kimi-VL-A3B-Instruct [35]	25.6	28.9	66.3	4.0	20.4	15.0	18.9	1.69	1.53	1.72	1.81	2.67	3.57	2.40	2.05
Kimi-VL-A3B-Think [35]	26.7	27.0	51.6	7.4	24.4	25.4	24.4	1.61	1.39	1.68	1.75	2.61	3.35	2.34	2.15
InternVL3-8B [55]	31.3	39.6	53.5	4.0	30.3	24.1	36.2	1.96	1.88	1.92	2.09	2.75	3.52	2.53	2.21
InternVL3-14B [55]	35.7	42.5	62.0	4.9	27.4	23.6	54.1	2.08	2.01	2.01	2.22	2.86	3.67	2.62	2.29
InternVL3-78B [55]	39.3	43.5	72.5	5.3	29.4	28.7	56.1	2.79	2.74	2.75	2.89	2.90	3.64	2.64	2.43
Qwen2.5-VL-3B [3]	26.2	30.8	56.1	5.7	29.9	21.2	13.8	1.00	0.83	1.05	1.12	2.15	2.98	1.77	1.71
Qwen2.5-VL-7B [3]	31.2	28.3	66.6	4.7	34.2	29.3	23.9	1.75	1.69	1.71	1.85	1.95	2.53	1.83	1.49
Qwen2.5-VL-32B [3]	35.3	36.7	54.7	11.6	33.2	**30.9**	44.8	1.55	1.42	1.52	1.72	2.96	3.63	2.71	2.55
Qwen2.5-VL-72B [3]	40.5	47.0	74.8	6.8	**34.8**	28.9	50.8	2.01	1.99	2.00	2.05	2.92	3.79	2.70	2.27
GeoChat-7B [14]	10.7	6.1	-	-	-	-	15.3	-	-	-	-	-	-	-	-
TeoChat-7B [13]	23.0	6.9	64.9	2.0	22.5	23.3	18.2	1.77	1.61	1.74	1.96	1.95	2.59	1.77	1.49
EarthDial-4B [34]	22.9	10.6	58.1	3.2	30.2	20.8	14.5	1.53	1.22	1.64	1.73	2.42	3.21	2.08	1.98
• Commercial models															
GPT-4o [12]	39.3	49.4	80.5	10.6	24.2	21.4	49.8	2.27	2.25	2.28	2.28	3.19	3.92	2.95	2.69
GPT-4.1 [12]	**42.3**	**52.4**	79.6	7.2	25.5	25.0	**64.0**	2.57	2.60	2.58	2.54	3.14	**3.94**	2.93	2.56
• Fine-tuned models															
Qwen2.5-VL-7B [3]	40.4	37.7	**83.6**	21.5	34.3	29.4	36.2	**3.90**	**3.76**	**3.53**	**4.41**	3.11	3.73	2.88	2.73
Δ	$\uparrow$9.2	$\uparrow$9.4	$\uparrow$17.0	$\uparrow$16.8	$\uparrow$0.1	$\uparrow$0.1	$\uparrow$12.3	$\uparrow$2.15	$\uparrow$2.07	$\uparrow$1.82	$\uparrow$2.56	$\uparrow$1.16	$\uparrow$1.20	$\uparrow$1.05	$\uparrow$1.24
InternVL3-8B [55]	41.7	42.6	79.3	**23.9**	29.1	24.9	50.6	3.83	3.69	3.49	4.32	**3.31**	3.92	**3.10**	**2.90**
Δ	$\uparrow$10.4	$\uparrow$3.0	$\uparrow$25.8	$\uparrow$19.9	$\downarrow$-1.2	$\uparrow$0.8	$\uparrow$14.4	$\uparrow$1.87	$\uparrow$1.81	$\uparrow$1.57	$\uparrow$2.23	$\uparrow$0.56	$\uparrow$0.40	$\uparrow$0.57	$\uparrow$0.69

Larger VLMs achieve higher performances. By scaling up LLMs, InternVL3 and Qwen2.5-VL series demonstrate consistent trends that larger LLMs achieve superior performances, confirming established scaling laws observed in general-domain applications. The commercial models, i.e., GPT-4o and GPT-4.1, showcase competitive performances across all tasks due to their massive corpus.

Remote sensing VLMs still struggle with disaster tasks. Despite being specifically trained on aerial and satellite imagery, existing remote sensing VLMs exhibit feature representations that inadequately transfer to the unique characteristics of disaster scenarios. DisasterM3 narrows the domain gap by providing specialized disaster-focused vision-language data for Earth observation applications.

Fine-tuned models improve comprehensively. By fine-tuning on DisasterM3 `Instruct` set, the performances of Qwen2.5-VL and InternVL3 have been significantly improved, narrowing the domain gap. Disaster-specific terminology integration during training significantly enhances report generation quality, resulting in more reasonable and professional reports. However, for building damage counting (BDC) task, the fine-tuned InternVL3 exhibits unexpected performance degradation due to overfitting, and we perform detailed analysis in §4.2. In the future, object sensitive module [54] and numerical enhanced optimization [41] could be explored for model development.

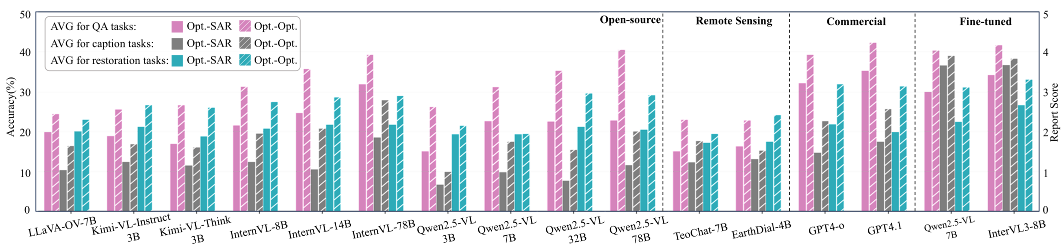

Figure 7: Benchmarking results of VLMs on DisasterM3 `Bench` set with optical-SAR setting.

Underrepresentation for SAR images. Disasters are often accompanied by extreme weather, with clouds and rain blocking optical sensors. In this case, the active imaging method SAR can penetrate clouds and fog to obtain accurate surface information. Fig. 7 shows the VLMs' performances evaluated on paired optical-SAR images. Due to the reduced semantics compared to optical imagery and underrepresentation in generic VLM, the performance using post-SAR images yielded substantially diminished performance across all evaluation tasks. In this scenario, as demonstrated by the Qwen2.5-VL series, increasing LLM size fails to yield stable improvements. The fine-tuned models alleviate the multi-sensor feature alignment, significantly improving the performances on all tasks. Since there still exists a huge gap compared to the optical-optical setting, more multi-modal pretraining [10] and alignment strategies could be investigated [40] for further improvement.

Table 3: Benchmarking results of referring segmentation VLMs on DisasterM3 `Bench` set. The accuracies across damage-levels (buildings and roads) are combined for simplification.

Model	Optical-Optical (%)					Optical-SAR (%)			
	mIoU	cIoU	Road	Building	Other	mIoU	cIoU	Road	Building
• Open-source models									
PSALM-1.3B [52]	9.7	6.3	2.6	10.2	16.3	8.1	8.8	5.1	11.1
HyperSeg-3B [44]	16.6	14.5	7.5	17.0	25.4	8.8	10.3	4.5	13.1
LISA-7B [15]	27.5	22.1	11.9	25.0	45.6	10.9	12.7	6.0	15.7
GeoPixel-7B [32]	14.3	14.2	8.5	18.1	16.2	4.0	5.1	1.8	6.2
• Fine-tuned models									
LISA-7B [15]	<u>44.8</u>	<u>43.7</u>	<u>27.6</u>	<u>41.2</u>	<u>65.5</u>	<u>28.2</u>	<u>29.6</u>	<u>21.5</u>	<u>34.9</u>
Δ	↑17.3	↑21.6	↑15.7	↑16.2	↑19.9	↑17.3	↑16.9	↑15.5	↑19.2
PSALM-1.3B [52]	**50.5**	**44.5**	**30.5**	**49.1**	**71.9**	**31.8**	**35.2**	**24.3**	**39.3**
Δ	↑40.8	↑38.2	↑27.9	↑38.9	↑55.6	↑23.7	↑26.4	↑19.2	↑28.2

Mask token matters in disaster referring segmentation. Tab. 3 shows the compared results of referring segmentation models with multi-sensor settings. After fine-tuning, LISA and PSALM have achieved significant gains in two settings with the injection of disaster reasoning knowledge during the training. It is notable that PSALM exceeds LISA with much smaller parameters. We attribute this to a more robust mask token representation in PSALM. Unlike LISA, which relies on a single fixed mask token for decoding, PSALM adopts a Mask2Former approach that generates comprehensive mask proposals through multiple mask tokens. We empirically set the number of mask tokens to 100 and observed that performance stabilizes when using more than 50 tokens. Disaster scene referring segmentation usually encompasses diverse objects at varying scales, necessitating robust mask token representations facilitated by LLMs.

4.2 Detailed Analysis

Performance variation across disaster categories. VLM performance exhibits variation across disaster types due to differing disaster causal factors and prone environments. As shown in Tab. 4, all methods demonstrate higher performance on landslide events while achieving notably lower metrics on earthquake, tornado, and explosion scenarios. This is because landslides often occur in rural mountainous regions, presenting simpler scenes with fewer objects. In contrast, the others primarily originate from highly developed urban areas, representing substantially more complex scenes. Due to multi-disaster events, the DisasterM3 dataset could measure VLMs comprehensively with unified metrics for multiple vision-language tasks.

Table 4: Performance variation across disaster categories. Accuracy (%) is calculated for each disaster category across all QA tasks.

Method	AVG	Landslide	Earthquake	Tornado	Conflict	Fire	Explosion	Tsunami	Hurricane	Volcano	Flooding
LLaVA-OV [17]	21.2	23.6	17.1	19.2	22.8	25.4	18.1	19.5	23.2	23.5	19.9
Kimi-VL-A3B-Instruct [35]	19.9	22.2	17.5	20.1	16.3	26.3	13.2	19.2	21.9	23.5	18.4
Kimi-VL-A3B-Think [35]	22.0	26.4	19.6	21.0	17.4	26.9	16.9	22.6	21.8	25.0	22.3
InternVL3-8B [55]	27.5	41.7	22.2	24.4	21.7	33.0	20.9	28.0	27.3	28.8	27.0
InternVL3-14B [55]	30.0	48.6	22.7	26.6	27.2	33.7	21.1	27.3	29.9	31.3	31.5
InternVL3-78B [55]	31.8	48.6	26.3	27.9	25.0	37.1	25.6	32.0	31.7	32.9	30.9
Qwen2.5-VL-3B [3]	24.5	26.4	19.5	21.9	27.5	32.1	17.9	24.2	24.2	29.6	21.8
Qwen2.5-VL-7B [3]	25.6	34.3	21.0	24.9	17.3	33.7	16.7	28.3	27.8	25.6	25.9
Qwen2.5-VL-32B [3]	31.0	50.0	26.4	27.1	26.6	35.7	23.9	32.8	29.4	30.8	27.7
Qwen2.5-VL-72B [3]	31.8	47.2	25.0	**31.1**	19.0	39.0	24.0	**33.4**	34.0	33.9	31.2
GPT-4o [12]	30.7	52.8	24.8	25.7	19.6	33.7	25.6	28.0	29.7	35.1	32.0
GPT-4.1 [12]	32.4	51.4	**26.9**	26.7	21.7	35.2	**27.7**	28.5	33.4	35.8	**36.5**
• Fine-tuned models											
Qwen2.5-VL-7B [3]	32.9	41.7	26.5	30.7	**27.7**	**40.3**	22.3	33.0	34.0	**41.8**	31.1
Δ	↑7.3	↑7.4	↑5.5	↑5.8	↑10.4	↑6.6	↑5.6	↑4.7	↑6.2	↑16.2	↑5.2
InternVL3-8B [55]	**34.7**	**56.9**	26.0	**31.1**	26.1	**40.3**	27.4	33.1	**34.9**	39.4	32.2
Δ	↑7.2	↑15.2	↑3.8	↑6.7	↑4.4	↑7.3	↑6.5	↑5.1	↑7.6	↑10.6	↑5.2

Performance biases in VLMs for damage counting. Remote sensing imagery typically encompasses numerous objects exhibiting diverse scales and morphologies, with counting challenges becoming particularly acute when conducting fine-grained damage assessment. Fig. 8 illustrates building damage assessment accuracy as a function of building density within analyzed scenes. For InternVL series models, performance initially declines and then improves as density increases. For peripheral ranges (<50 or $\geq$200 buildings), these models demonstrate higher confidence and accuracy. In contrast, GPT-4 models exhibit a clear inverse relationship between building density and accuracy. The fine-tuned InternVL3-8B exhibits substantial improvement in low-density scenes (<50 buildings) but notable degradation in other ranges, revealing an overfitting dilemma. Different VLMs have different biases in the damage assessment task. In the future, we can integrate pixel-level semantics provided by the DisasterM3 dataset to alleviate the overfitting risk.

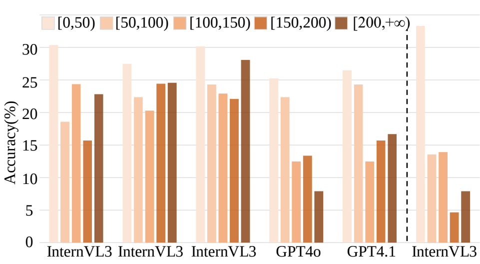

Figure 8: Counting acc v.s. Object density.

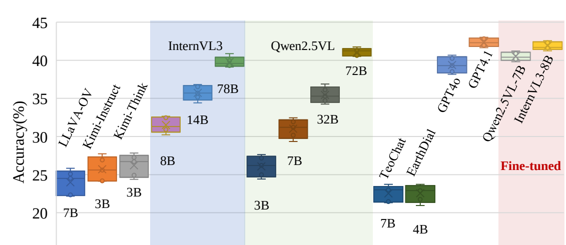

Figure 9: Accuracy variation with different prompts.

Effects of different prompts. As shown in Fig. 9, we evaluated the robustness of VLMs with five different prompts, where quartiles, ranges of accuracies are plotted. Due to limited LLM capabilities, LLaVA-OV, TeoChat, EarthDial, and Kimi models exhibit higher sensitivity to prompt variations.

Besides, InternVL3 and Qwen2.5-VL series models show similar patterns wherein larger LLMs display enhanced stability. In comparison to GPT-4o, GPT-4.1 achieves superior performance with notably improved consistency. Following enrichment with the disaster-specific corpus from the DisasterM3 dataset, the fine-tuned Qwen2.5-VL-7B and InternVL3-8B model demonstrate good stability to prompt variations.

5 Limitations and Future Directions

While DisasterM3 represents a significant step forward in disaster-oriented vision-language research, we acknowledge several limitations that open avenues for future work. **1) Multi-resolution generalization**: Our standardization to 0.8 m resolution ensures controlled experimentation but limits evaluation of model robustness to diverse spatial resolutions encountered in operational settings. Future work should incorporate multi-resolution imagery from platforms like Sentinel-2 (10m) and Landsat (30m), leveraging our existing annotations through geo-registration. **2) Enhanced sensor diversity**: Although we include both optical and SAR imagery, our SAR data is limited to single polarization. Integrating multi-polarization data (e.g., Sentinel-1's VV+VH) would provide richer scattering information about debris orientation and surface characteristics, enabling more comprehensive damage assessment. **3) Cross-sensor performance gap**: The significant performance degradation on SAR imagery highlights the need for advanced multi-modal pretraining and cross-modal alignment strategies to better bridge the optical-SAR domain gap. **4) Counting task optimization**: To address overfitting in damage object counting, promising directions include object-sensitive encoders (e.g., DINOv2), numerical difference loss, synthetic data generation via diffusion models for high-density scenarios, and knowledge distillation strategies. **5) Living benchmark commitment**: We will maintain DisasterM3 as an evolving resource by regularly incorporating new disaster events from the Maxar Open Data Program, ensuring continued relevance and growth in geographic and temporal coverage for the disaster response community.

6 Conclusion

Inspired by the rapid development of generic VLMs, the remote sensing vision-language datasets and methods have also been gradually explored. To promote interactive AI disaster response, we propose DisasterM3, a multi-hazard, multi-sensor, and multi-task remote sensing dataset for vision-language understanding. DisasterM3 includes 26,988 bi-temporal images and 123k instruction pairs, 36 disaster events across 5 continents. The comprehensive benchmarking of 14 advanced VLMs evaluate both their capabilities and inherent limitations in disaster contexts. Additionally, through fine-tuning four VLMs with the disaster-specific corpus from DisasterM3, we demonstrate substantial performance enhancements across all evaluation tasks. We believe the proposed dataset and baselines will help bridge the gap in VLM-based disaster applications within Earth vision.

Acknowledgments

This work was supported in part by the Council for Science, Technology and Innovation (CSTI) and the Cross-ministerial Strategic Innovation Promotion Program (SIP) "Development of a Resilient Smart Network System against Natural Disasters" (funding agency: NIED), KAKENHI (25K03145) as well as the NVIDIA Academic Grant Program. This work used computational resources Miyabi supercomputer provided by The University of Tokyo through Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures and High Performance Computing Infrastructure in Japan (Project ID: jh250017). Weihao Xuan is supported by RIKEN Junior Research Associate (JRA) Program. We also thank Ritwik Gupta for sharing the valuable xBD dataset and for his expertise in disaster response guidance.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Anthropic. Model card: Claude 3. Technical report, Anthropic, 2024.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Capella Space. Earth observation gallery, 2025. https://www.capellaspace.com/earth-observation/gallery.

[5] Hongruixuan Chen, Jian Song, Olivier Dietrich, Clifford Broni-Bediako, Weihao Xuan, Junjue Wang, Xinlei Shao, Yimin Wei, Junshi Xia, Cuiling Lan, et al. Bright: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *arXiv preprint arXiv:2501.06019*, 2025.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[7] Radiological Emergency. Federal emergency management agency, 2002.

[8] Elizabeth Frankenberg, Cecep Sumantri, and Duncan Thomas. Effects of a natural disaster on mortality risks over the longer term. *Nature sustainability*, 3(8):614–619, 2020.

[9] 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. *arXiv preprint arXiv:1911.09296*, 2019.

[10] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024.

[11] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023.

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[13] Jeremy Andrew Irvin, Emily Ruoyu Liu, Joyce Chuyi Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. *arXiv preprint arXiv:2410.06234*, 2024.

[14] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.

[15] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[17] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[18] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *NeurIPS*, 2024.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[20] Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26658–26668, 2024.

[21] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110, 2024.

[22] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

[23] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[24] Maxar Technologies. Open data program, 2025. https://www.maxar.com/open-data.

[25] Jane Palmer. The new science of volcanoes harnesses ai, satellites and gas sensors to forecast eruptions. *Nature*, 581(7808):256–260, 2020.

[26] Hongyi Pan, Diaa Badawi, Chang Chen, Adam Watts, Erdem Koyuncu, and Ahmet Enis Cetin. Deep neural network with walsh-hadamard transform layer for ember detection during a wildfire. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 257–266, 2022.

[27] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[28] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133, 2024.

[29] 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.

[30] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[31] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[32] Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad S. Khan, and Salman Khan. Geopixel : Pixel grounding large multimodal model in remote sensing. *ArXiv*, 2025.

[33] Dan H Shugar, Mylène Jacquemart, David Shean, Shashank Bhushan, Kavita Upadhyay, Ashim Sattar, Wolfgang Schwanghart, S McBride, M Van Wyk De Vries, M Mergili, et al. A massive rock and ice avalanche caused the 2021 disaster at chamoli, indian himalaya. *Science*, 373(6552):300–306, 2021.

[34] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[35] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[36] Isabelle Tingzon, Nuala Margaret Cowan, and Pierre Chrzanowski. Fusing vhr post-disaster aerial imagery and lidar data for roof classification in the caribbean. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3740–3747, 2023.

[37] Umbra. Open data program, 2025. https://umbra.space/open-data.

[38] United Nations Satellite Centre (UNOSAT). Emergency mapping products. `https://unosat.org/products`, 2025. Available at: `https://unosat.org/products`.

[39] Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? *arXiv preprint arXiv:2503.23771*, 2025.

[40] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[41] Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5481–5489, 2024.

[42] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[43] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023.

[44] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*, 2024.

[45] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[46] Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[47] Junxiang Xu, Divya Jayakumar Nair, and S Travis Waller. Implementing equitable wildfire response plans. *Science*, 388(6743):158–159, 2025.

[48] Weihao Xuan, Junjue Wang, Heli Qi, Zihang Chen, Zhuo Zheng, Yanfei Zhong, Junshi Xia, and Naoto Yokoya. DynamicVL: Benchmarking multimodal large language models for dynamic city understanding. *arXiv preprint arXiv:2505.21076*, 2025.

[49] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024.

[50] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, June 2024.

[51] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[52] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2025.

[53] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[54] Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Zheng Wang, Xian Zhong, and Shengfeng He. Zero-shot object counting with good exemplars. In *European Conference on Computer Vision*, pages 368–385. Springer, 2024.

[55] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim the contribution of the DisasterM3 dataset and conclude the main experimental results in the abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The InternVL3-8B model, following fine-tuning on our dataset, exhibits potential overfitting tendencies on damage counting tasks, as comprehensively analyzed in Section 4.2.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper focuses on the dataset and benchmark and does not include theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have detailed all experimental settings, evaluation metrics in this paper for reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset and code are provided here. We provided detailed instructions, such as prompt designing and fine-tuning details, in the Appendix §B.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the training and test settings in Section 4, and more details are in the Appendix §B.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We test the robustness of VLMs with different prompts and report the performance variations in Section 4.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have clarified the model training and testing sources. All experiments were conducted on 4 NVIDIA H100 GPUs with 96GB of memory.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] .

    Justification: The broader impacts are clarified in the Appendix §I.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the used datasets and code are open-source and properly cited.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use the GPT4-o to generate incorrect options and polish manual answers.