A Dataset Quality Control

To ensure the quality of dataset annotation, we construct a multi-step labeling and check framework in Fig. 10.

1) Annotator Training. All remote sensing and disaster experts undergo training based on guidelines established by the United Nations Institute for Training and Research (UNITAR) and the Federal Emergency Management Agency (FEMA), acquiring specialized knowledge of disaster-specific terminology, definitions, and assessment protocols.

2) First-round annotation. Following training, the qualified annotators are organized into three independent teams, each tasked with annotating a distinct subset of disaster samples during the initial assessment phase.

3) Cross validation. Following the initial assessment phase, we implemented a rigorous cross-validation protocol in which each team systematically reviewed the annotations produced by the other teams to ensure consistency and accuracy across the dataset. Samples identified as inconsistent or inadequate during the cross-validation process were flagged and returned to their original annotation team for comprehensive revision.

4) Expert verification. Team leaders subsequently performed quality assurance by randomly sampling 10-20% of the annotated data for verification, systematically identifying common patterns of error, recurring inconsistencies, and instance-specific issues requiring secondary revision. This iterative annotation-validation cycle (steps 2-4) was conducted multiple times until all samples met rigorous quality standards and achieved high inter-annotator agreement.

5) Comprehensive evaluation. Based on the DisasterM3 dataset, we conducted several statistical analyses, checking the outliers. In addition, we also used GPT-4.1 to evaluate the semantic consistency between multi-level questions for the same scene. Finally, we performed the preliminary experiments for validation.

The standard quality control framework strictly ensures the quality of data annotation. When a new disaster occurs, it is easy to extend new data using the proposed annotation pipeline and quality control framework.

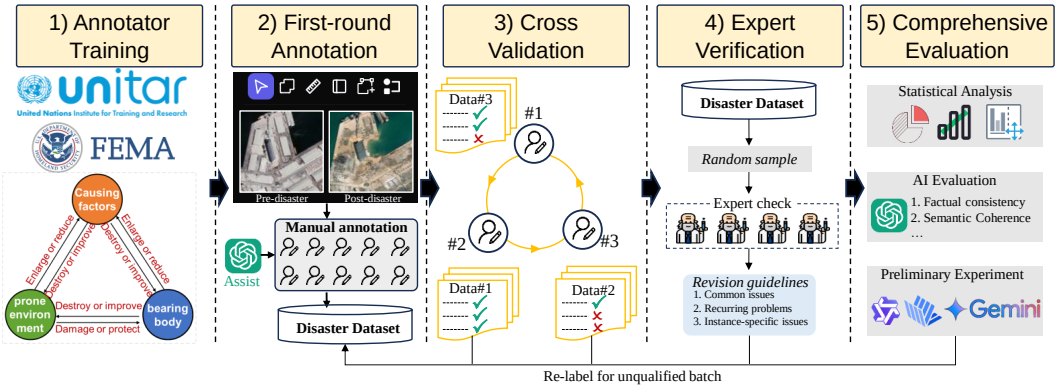

Figure 10: Dataset quality control framework includes five steps, ensuring the high-quality of dataset annotation.

A.1 Building damage-level definitions.

Following FEMA guidelines, we established clear building damage level criteria for annotator training using standardized definitions (Tab. 5) to enable annotators to develop robust visual feature recognition skills for accurate damage-level classification.

Table 5: Building damage categories and definitions.

Category	Definition
Background	All non-building pixels.
Intact	No visible signs of structural damage, water intrusion, shingle displacement, or burn marks.
Damaged	Partial structural damage to the building, such as missing roof members, visible cracks, or partial collapse of the wall/roof. Buildings may be partially burned, surrounded by water or mud, or affected by nearby volcanic flows.
Destroyed	Completely collapsed, burned, partially/completely covered by water/mud, or no longer present.

B Implementation details

B.1 Benchmark model settings.

We implement all open-sourced benchmark models using the vLLM toolkit [3]. We adopt each model's default input configurations for benchmarking.

For referring segmentation evaluation, we utilize four state-of-the-art models with their default source code configurations. LISA employs a LLaVA-based architecture with CLIP VIT-L/14 as the vision encoder (224×224 resolution) and LLaMA-2-7B as the language model backbone.

PSALM adopts a Mask2Former-based architecture that unifies multiple segmentation tasks through a flexible input schema. PSALM adopts Swin-B visual backbone and Phi-1.5 1.3B as a language model. Before its Mask2Former-style query decoder, 100 learnable mask tokens are introduced to unify the multi-task segmentation input mode.

HyperSeg represents the first VLLM-based universal segmentation model, integrating a fine-grained pyramid visual encoder, a lightweight Mipha language model, and a Mask2Former predictor.

GeoPixel is specifically designed for remote sensing imagery, featuring an adaptive image divider that partitions inputs into local and global regions to handle resolutions up to 4K in any aspect ratio. The architecture comprises scaled CLIP ViT-L/14, InternLM2 language model with partial LoRA adaptation, and SAM-2 integrated pixel decoder.

B.2 Model fine-tuning details

B.3 InternVL3 and Qwen2.5-VL

We employ a standard Low-Rank Adaptation (LoRA) fine-tuning strategy to optimize the Large Language Model (LLM) component of InternVL3 and Qwen2.5-VL. During this process, we freeze the vision encoder and fine-tune only the LLM. We train the model using the question-answering samples from our DisasterM3, configuring the LoRA module with a rank of 64, alpha of 16, and dropout rate of 0.05. The training is conducted on 4 H100 GPUs with a global batch size of 256 for one epoch. We use the AdamW optimizer with a learning rate of 2×10^{-4}, setting β_1=0.9 and β_2=0.95, and apply a cosine learning rate scheduler.

B.4 LISA and PSALM

LISA Fine-Tuning: We conducted LoRA fine-tuning based on the LISA 7B pre-trained model, utilizing segmented instruction-tuning data from DisasterM3. Since the LoRA parameters of the LISA pre-trained model had already been merged with the base model, our LoRA parameters were randomly initialized. Throughout this process, we adopted LISA's original training configuration, specifically configuring the LoRA module with a rank of 8, alpha of 16, and a dropout rate of 0.05. We employed the AdamW optimizer with a learning rate of 3×10^{-4} and implemented a cosine learning rate scheduler. The training was conducted for 1,000 steps to prevent overfitting to our dataset. We utilized a batch size of 64 with gradient accumulation steps of 8, performing the fine-tuning process across 4 H100 GPUs.

PSALM Fine-Tuning: We employed the PSALM Phi1.5 1.3B version as our base model and followed its original training configuration. The model was trained for 10 epochs using a batch size of 64, with fine-tuning conducted on 4 H100 GPUs. In contrast to the LISA fine-tuning approach, we adopted PSALM's native configuration by keeping the LLM parameters unfrozen and performing direct fine-tuning. This methodology ensures optimal performance within the PSALM framework architecture.

C Design of Instruction prompts

C.1 Instruction prompts

Tab. 6 presents a comprehensive framework of instruction prompts designed for the DisasterM3 `Bench` set, encompassing six distinct disaster analysis tasks and two comprehensive reports. The instruction templates follow a structured approach, where each task requires analysis of pre-disaster and post-disaster satellite imagery pairs. For classification-based tasks (DSR and BBR), the prompts elicit multi-label responses formatted as comma-separated capital letters. Single-choice tasks (DTR, BDC, and DRE) require simplified single-letter responses. The ORR task uniquely focuses on spatial relationship analysis using a single image with highlighted objects. Beyond these discrete tasks, two complex reports are introduced: Disaster Caption, which demands a comprehensive multi-category impact assessment structured across six environmental domains (disaster type,

[3] https://github.com/vllm-project/vllm

buildings, roads, vegetation, water bodies, agriculture, and an overall conclusion); and Restoration Advice, which requires actionable recovery recommendations segmented into immediate and long-term strategies. This instruction design systematically evaluates a model's capacity to process multi-temporal disaster imagery while enforcing strict output formatting requirements that facilitate automated performance evaluation.

Table 6: The instruction prompts of DisasterM3 `Bench` set for different tasks. DSR-disaster scene recognition, BBR-Bearing-body damage recognition, DTR-damage type recognition, DBC-damage building counting, DRE-damage road estimation, ORR-object relational reasoning.

Instruction Templates	Task	Question Prompts
Analyze both the pre-disaster and post-disaster images to answer the following question. Choose the best option(s) from the candidate options provided. pre-disaster image: <image> post-disaster image: <image> Question: <question_prompt> Options: <options> Your task is to respond with ONLY the capital letters of the correct options, separated by a comma and a space (e.g., C, D, H). Do not include any explanation.	**DSR**	· Can you identify and categorize the different types of land use visible in this pre-disaster satellite image? · Identify the main land-use types present before the disaster. · Classify the land-use zones in this pre-disaster scene. · What land-use patterns appear in this pre-disaster imagery? · What land-use categories are visible in this pre-disaster image?
	BBR	· Which key objects show visible impact from this disaster event? · Identify the primary objects compromised in this disaster scene. · What essential land-cover objects appear damaged in this disaster zone? · What categories of objects have sustained damage in the affected area? · Which critical objects exhibit disaster-related damage?
Analyze both the pre-disaster and post-disaster images to answer the following question. Choose the best option from the candidate options provided. pre-disaster image: <image> post-disaster image: <image> Question: <question_prompt> Options: <options> Your task is to respond with ONLY the capital letter of the correct option (e.g., C). Do not include any explanation or other text.	**DTR**	· What disaster has happened in this area? · Identify the disaster that has impacted this location. · What disaster event has taken place in this area? · What type of disaster occurred in this region? · What kind of calamity has this area experienced?
	BDC	· What is the total number of completely destroyed buildings? · Count the buildings that were totally destroyed. · What's the count of buildings that were utterly demolished? · How many buildings were totally destroyed? · What is the total count of buildings that were fully devastated?
	DRE	· What percentage of the entire image is occupied by flooded roads? · Calculate what fraction of the whole image is taken up by submerged roads. · What proportion of the total image area consists of roads covered by flood water? · What is the ratio of flooded road area to the entire image? · What is the proportion of the complete image that consists of flooded roads?
Analyze the image to answer the following question. Choose the best option from the candidate options provided. Image: <image> Question: <question_prompt> Options: <options> Your task is to respond with ONLY the capital letter of the correct option (e.g., C). Do not include any explanation or other text.	**ORR**	· Explain how object in red box spatially relates to object in blue box. · Describe the spatial relationship between object in red box and object in blue box. · Explain how object in red box is spatially positioned relative to object in blue box. · Characterize the positional relationship that exists between object in red box and object in blue box. · How does object in red box relate spatially to object in blue box?
Disaster Caption		
Your TASK is to analyze the provided pair of pre-disaster and post-disaster remote sensing images. You will act as a remote sensing analyst to identify the type of disaster and assess its impact on both built and natural environments across five specific categories. pre-disaster image: <image> post-disaster image: <image> Your analysis must be formatted as follows: DISASTER: [the name of the disaster] BUILDING: [describe impacts on buildings] ROAD: [describe impacts on road networks] VEGETATION: [describe impacts on natural, unmanaged vegetation cover] WATER_BODY: [describe changes to water bodies] AGRICULTURE: [describe impacts on managed agricultural land] CONCLUSION: [provide a concise 1-2 sentence summary synthesizing the overall disaster impacts observed across the categories.]		
Restoration Advice		
Your TASK is to generate concise and integrated recovery recommendations for the affected area based on the provided pre-disaster and post-disaster remote sensing images. Aspects to focus on include infrastructure restoration, housing reconstruction, and ecological and geological environment restoration. pre-disaster image: <image> post-disaster image: <image> Based on your analysis of the images: 1. First determine if recovery actions are necessary. If no significant damage or impact is observed, clearly state no recovery recommendations due to no discernible impact. 2. If recovery is needed, provide recommendations in the following format: IMMEDIATE_RECOVERY: [Provide an integrated paragraph within 50 words describing immediate recovery actions. Create a flowing narrative.] LONG_TERM_RECOVERY: [Provide an integrated paragraph within 50 words describing long-term recovery strategies. Create a flowing narrative.] Ensure your recommendations are realistic, feasible, and properly prioritized based on the visible damage in the images.		

C.2 GPT-based Evaluation Rubric and Prompts

Tab 7 presents the evaluation frameworks designed for assessing two complex tasks in the DisasterM3 Bench dataset. For the Disaster Caption task, we developed a three-dimensional evaluation criteria: Damage Assessment Precision evaluates the accuracy between predicted descriptions and actual damage situation; Damage detail recall measures the completeness of disaster captions ; and Factual correctness evaluates fabricated content in predictions that does not exist in ground truth annotations or would not be visible in the images.

The Disaster Restoration Advice task is evaluated through four dimensions: Recovery Necessity Recognition judges the correct acknowledgment of whether recovery actions are necessary; Action Priority Precision measures the alignment of suggested actions with reference plan priorities; Strategic Completeness assesses the coverage of key recovery elements; and Implementation Feasibility evaluates the practicality and applicability of the recommendations. Both task evaluations employ a 0-5 integer scoring system, requiring evaluators to provide brief explanations to justify their scores, ensuring transparency and consistency in the assessment process. This structured evaluation framework provides comprehensive, fine-grained quantitative metrics for the performance of large vision-language models in disaster analysis tasks.

Table 7: Evaluation prompts for GPT-4.1: Disaster Caption and Restoration Advice

Disaster Caption Evaluation
You are an advanced intelligent chatbot specialized in evaluating the accuracy of disaster scene captions that compare pre-disaster and post-disaster images. Your primary task is to meticulously compare the predicted caption with the ground truth caption and assess their factual consistency. To accomplish this, you will evaluate the captions across four key dimensions: 1. **Damage Assessment Precision:** Evaluate how accurately the elements mentioned in the predicted caption match the actual damage described in the ground truth caption. This measures whether the predicted details are correct (without considering comprehensiveness). 2. **Damage Detail Recall:** Assess how completely the predicted caption captures all the damage elements mentioned in the ground truth caption. This measures whether the prediction includes all relevant damage information from the ground truth. 3. **Factual Correctness:** Evaluate the absence of hallucinated content. Higher scores indicate fewer or no hallucinations, while lower scores indicate more hallucinations (facts, elements, or interpretations that do not exist in the ground truth caption or would not be visible in the images). Please assign a score for each of these three dimensions, using an integer from 0 to 5, where 5 indicates perfect performance and 0 signifies poor performance. Accompany your assessments with brief explanations to clarify your scoring rationale.
Disaster Restoration Advice Evaluation
You are an advanced intelligent evaluator specialized in assessing disaster recovery plans that compare recommended immediate and long-term recovery strategies following disasters. Your primary task is to meticulously compare the predicted recovery plan with the ground truth recovery plan and assess their factual consistency and strategic alignment. To accomplish this, you will evaluate the recovery plans across four key dimensions: 1. **Recovery Necessity Recognition:** Assess whether the predicted plan correctly recognizes if recovery actions are necessary. If the ground truth indicates no recovery is needed (e.g., "no discernible impact detected"), the prediction should similarly acknowledge this. Conversely, if the ground truth outlines necessary recovery actions, the prediction should not minimize or overlook the need for recovery. 2. **Action Priority Precision:** Evaluate how accurately the specific recovery actions mentioned in the predicted plan match the priorities described in the ground truth plan. This measures whether the predicted recovery actions are correct (without considering comprehensiveness). If no recovery is needed according to both plans, award full points. 3. **Strategic Completeness:** Assess how completely the predicted plan captures all the essential recovery elements mentioned in the ground truth plan. This measures whether the prediction includes all relevant recovery strategies from the ground truth. If no recovery is needed according to both plans, award full points. 4. **Implementation Feasibility:** Evaluate the practicality and absence of unrealistic recommendations. Higher scores indicate realistic, implementable recovery actions, while lower scores indicate impractical suggestions or approaches that would be ineffective in the described disaster context. If no recovery is needed according to both plans, award full points. Please assign a score for each of these four dimensions, using an integer from 0 to 5, where 5 indicates perfect performance and 0 signifies poor performance. Accompany your assessments with brief explanations to clarify your scoring rationale.

D Experimental results on Optical-SAR setting

Tab. 8 presents comprehensive evaluation results of various VLMs on our DisasterM3 Bench with Optical-SAR setting. The evaluation encompasses both multiple-choice tasks (measured by accuracy percentage) and open-ended generation tasks (scored by GPT-4.1 on a 5-point scale). Several key observations emerge from the performance analysis across different model categories and task types. Commercial models demonstrate superior performance, with GPT-4.1 achieving the highest overall accuracy of 35.2%, followed by GPT-4o at 32.1%. Among open-source models, InternVL3-78B leads with 31.8% accuracy, significantly outperforming other models in its category. The fine-tuned models show competitive results, with InternVL3-8B reaching 34.1% after domain-specific training. As for multiple-choice tasks, performance varies significantly across different recognition and reasoning tasks. Disaster Type Recognition (DTR) proves most tractable, with top-performing models achieving over 70% accuracy (GPT-4o: 73.1%, GPT-4.1: 71.6%, InternVL3-8B fine-tuned: 73.1%). Object Relational Reasoning (ORR) also shows reasonable performance, with GPT-4.1 reaching 49.4%. However, Bearing-Body Damage recognition (BBR) remains extremely challenging, with the best model (Qwen2.5-VL-72B) achieving only 22.1% accuracy. This because This is because SAR contains limited information and cannot recognize the natural objects.

Open-ended tasks reveal interesting patterns in model capabilities. For disaster caption, fine-tuned models dramatically outperform their base versions, with fine-tuned Qwen2.5-VL-7B achieving 3.65 average score compared to 0.98 for the base model—representing a 3.7× improvement. Among caption sub-metrics, Factual Correctness (FC) consistently scores highest across models, while Damage Assessment Precision (DAP) and Damage Detail Recall (DDR) show more modest performance, suggesting models struggle with precise damage quantification and comprehensive detail extraction. Recovery Necessity (RN) scores are consistently higher than Action Priority Precision (APP) and Strategic Completeness (SC) across all models. This pattern indicates that while models can identify areas requiring restoration, they struggle with prioritizing actions and providing comprehensive strategic guidance. Commercial models maintain relatively balanced performance across all three restoration metrics, while open-source models show more variable performance.

Table 8: Benchmarking various VLMs on DisasterM3 Bench set with Optical-SAR setting.

Method	Accuracy (%)						Disaster Caption				Restoration Advice			
	AVG	DTR	BBR	BDC	DRE	ORR	AVG	DAP	DDR	FC	AVG	RN	APP	SC
Random Guess	-	*20*	-	*20*	*20*	*20*	-	-	-	-	-	-	-	-
• Open-source models														
LLaVA-OV-7B [17]	19.8	37.3	3.4	22.2	19.4	16.9	1.03	0.84	0.78	1.47	2.00	2.56	1.81	1.63
Kimi-VL-A3B-Instruct [35]	18.9	58.2	4.5	15.1	7.4	9.4	1.24	1.09	1.17	1.47	2.79	2.70	1.89	1.78
Kimi-VL-A3B-Think [35]	16.9	34.3	7.6	17.7	12.9	11.9	1.15	0.96	1.10	1.39	2.22	2.35	1.71	1.59
InternVL3-8B [53]	21.5	32.8	7.3	20.7	18.4	28.1	1.24	1.08	1.02	1.62	2.07	2.55	1.90	1.75
InternVL3-14B [53]	24.6	32.8	7.6	22.5	17.7	42.5	1.05	0.86	0.82	1.46	2.17	2.67	2.01	1.84
InternVL3-78B [53]	31.8	65.7	11.2	26.2	21.6	34.4	1.85	1.73	1.66	2.17	2.17	2.59	1.97	1.96
Qwen2.5-VL-3B [3]	15.0	23.9	7.3	23.3	13.9	6.9	0.67	0.55	0.62	0.84	1.93	2.55	1.65	1.58
Qwen2.5-VL-7B [3]	22.6	62.7	8.4	16.9	11.9	13.1	0.98	0.86	0.90	1.19	1.93	2.41	1.85	1.54
Qwen2.5-VL-32B [3]	22.5	37.3	11.8	20.3	14.5	28.7	0.77	0.56	0.60	1.14	2.12	2.58	1.90	1.89
Qwen2.5-VL-72B [3]	22.8	40.3	22.1	14.6	10.0	26.9	1.16	1.02	1.11	1.35	2.05	2.53	1.87	1.74
TeoChat [13]	15.0	29.9	4.5	18.4	9.4	13.1	1.23	1.08	1.09	1.51	1.72	2.20	1.58	1.38
EarthDial [34]	16.3	30.7	6.8	19.5	10.2	14.3	1.31	1.31	1.37	1.25	1.74	2.31	1.47	1.44
• Commercial models														
GPT-4o [12]	32.1	73.1	17.4	20.6	10.0	39.4	1.47	1.35	1.33	1.73	2.19	2.55	1.99	2.02
GPT-4.1 [12]	35.2	71.6	17.6	21.4	15.8	49.4	1.74	1.68	1.63	1.92	1.98	2.37	1.82	1.76
• Fine-tuned models														
Qwen2.5-VL-7B [3]	29.9	64.2	21.0	29.4	13.9	21.2	3.65	3.38	3.31	4.45	2.25	2.66	2.04	2.04
InternVL3-8B [53]	34.1	73.1	18.8	23.6	18.7	36.2	3.66	3.38	3.10	4.50	2.66	2.97	2.50	2.52

E Experimental results on numerical tasks

Because numerical tasks require more natural responses, we assessed VLM performance using Root Mean Square Error (RMSE) as the evaluation metric for Building Damage Counting (BDC) and Damage Road Estimation (DRE) tasks. RMSE quantifies the deviation between predicted values and ground truth annotations, and lower RMSE values indicate better counting accuracies. The comparative results between open-ended (RMSE) and multiple-choice questions (OA) are as follows:

Table 9: Results on BDC and DRE. Lower is better for RMSE; higher is better for OA (%).

Method	↓RMSE		↑OA (%)	
	BDC	DRE	BDC	DRE
LLaVA-OV	114.32	10.37	26.4	24.2
InternVL3-8B	86.93	10.17	30.3	24.1
InternVL3-14B	102.03	12.11	27.4	23.6
InternVL3-78B	105.96	9.53	29.4	28.7
Qwen2.5-VL-3B	95.04	17.86	29.9	21.2
Qwen2.5-VL-7B	69.66	4.27	34.2	29.3
Qwen2.5-VL-32B	76.61	3.91	33.2	30.9
Qwen2.5-VL-72B	53.83	7.83	34.8	28.9
GPT-4o	127.51	14.86	24.2	21.4
GPT-4.1	115.89	9.60	25.5	25.0
Qwen2.5-VL-7B (Fine-tune)	61.39	4.73	34.3	29.4
InternVL3-8B (Fine-tune)	108.88	10.18	29.1	24.9

The comparative performance demonstrates that models maintain consistent relative rankings across both evaluation formats, validating the robustness of our MCQ design.

F Scaling up LLMs on PSALM

To analyze the performances of PSALM with different LLMs, we have conducted additional experiments scaling up to larger language models using Qwen2.5-3B and Qwen2.5-7B on referring segmentation tasks. Fig. 11 shows three consistent trends. (1) **Fine-tuning is crucial.** The non–fine-tuned 1.3B model performs poorly (near-single digits cIoU), while fine-tuning on DisasterM3 yields a large jump. (2) **Bigger LLMs help.** Moving from 1.3B to 3B and 7B brings steady gains, with *Opt.-Opt.* improving by roughly ten points and *Opt.-SAR* by around five to seven points. (3) **Cross-sensor grounding is harder.** Despite overall improvements, the *Opt.-SAR* track remains notably below *Opt.-Opt.*, indicating a persistent modality gap.

We attribute the gains from scaling primarily to better linguistic disambiguation and more reliable phrase-to-region grounding, especially for complex spatial descriptions and multi-clause referring expressions. However, the cross-sensor gap suggests that scaling the LLM alone is insufficient when visual statistics shift (e.g., SAR backscatter vs. optical radiance). Bridging this gap likely requires sensor-aware visual encoders or adapters, SAR-specific augmentations, and additional paired/weakly paired multi-sensor supervision.

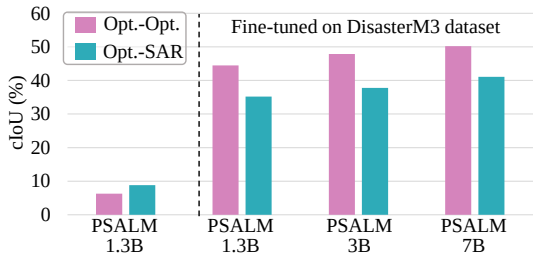

Figure 11: The compared results on PSALM with different LLMs for varied remote sensing sensors.

G Potential Geographic Bias

To assess potential geographic bias, we compare model performance on disasters originating in the United States versus those elsewhere (Tab. 10). Across all VLMs, results are well balanced between the two groups, and this holds regardless of whether the model is fine-tuned on our dataset.

We attribute this robustness to two factors: (1) our large-scale dataset provides substantial coverage of both US and non-US regions, and (2) disaster-related visual cues—such as structural damage, debris, and flooding—tend to be consistent across national boundaries. Together, these properties mitigate potential geographic bias.

Table 10: Results on US and no-US disasters.

Model	**US**	**No-US**
LLaVA-OV	24.84	24.15
Qwen2.5-VL-7B	31.18	31.23
Qwen2.5-VL-32B	35.42	35.17
Qwen2.5-VL-72B	40.70	40.28
InternVL3-8B	30.96	31.55
InternVL3-14B	35.94	35.38
InternVL3-78B	38.64	39.77
TEOChat	23.34	22.32
GPT-4.1	41.76	42.75
Qwen2.5-VL-7B (Fine-tune)	39.74	39.87
InternVL3-8B (Fine-tune)	41.88	41.43

H Visualizations on different disasters

In this section, we present representative visualizations across different disaster types from the DisasterM3 `Bench` set. As shown in Fig. 12, we demonstrate results for a flooding event, comparing model performance across multiple tasks: referring segmentation, disaster-bearing body recognition, damaged building counting, damaged road area estimation, and disaster captioning.

For the referring segmentation task with the prompt "Please help me identify and outline all areas inundated by floodwater after the disaster," generic VLMs including LISA, HyperSeg, and PSALM produce incorrect segmentations due to their lack of disaster-specific semantic understanding. Even GeoPixel, a specialized geospatial referring segmentation model, fails to accurately segment the flooded regions. However, after fine-tuning on our proposed DisasterM3 `Instruct` set, both LISA and PSALM successfully identify and segment the flooded areas, demonstrating the effectiveness of our disaster-specific dataset.

For the disaster bearing-body recognition task with the prompt "Which key objects show visible impact from this disaster event?", all baseline VLMs fail to identify the complete set of affected objects. In contrast, InternVL3-8B fine-tuned on DisasterM3 `Instruct` correctly identifies all impacted elements, providing the accurate answer "A, B, D."

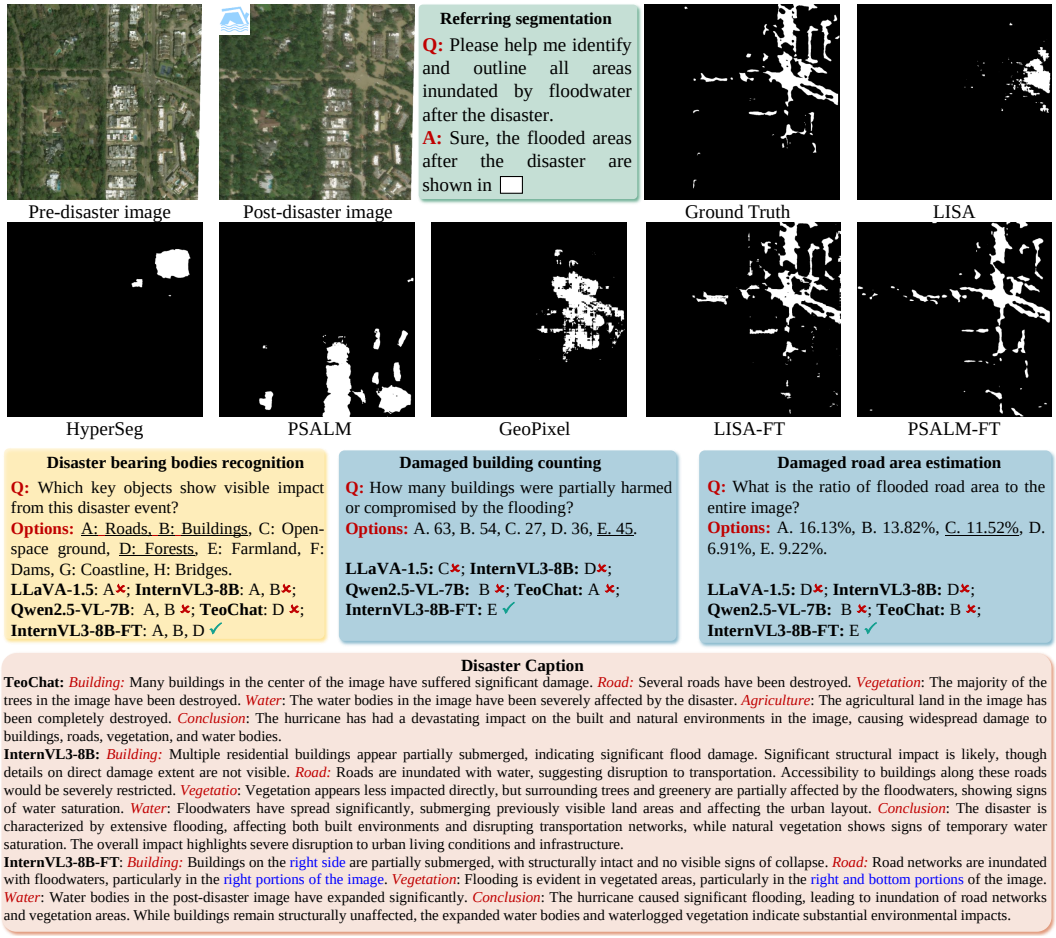

Figure 12: Visualization of compared predicted results for flooding disaster scene under the optical-optical setting.

Similarly, for the damaged building counting task with the prompt "How many buildings were partially harmed or compromised by the flooding?", all baseline methods fail to calculate the correct number of affected buildings due to their lack of disaster-specific terminology understanding. However, InternVL3-8B fine-tuned on DisasterM3 `Instruct` successfully identifies the accurate count of 45 buildings. For the damaged road area estimation task, we observe the same trend: baseline VLMs struggle to accurately quantify the affected road infrastructure, whereas the fine-tuned InternVL3-8B provides reliable area measurements.

For the disaster captioning task, we observe a clear performance hierarchy among the evaluated models. GeoChat produces vague, general descriptions and introduces factual errors, incorrectly describing agricultural damage in areas with no farmland present. Zero-shot InternVL3-8B shows significant improvement, generating detailed captions that largely correspond to the ground truth observations. Most notably, fine-tuning InternVL3-8B on our DisasterM3 `Instruct` dataset enables the model to incorporate precise spatial terminology, describing disaster impacts with location-specific references such as "right side" and "right and bottom portions."

As shown in Fig. 13, we demonstrate results for an earthquake event, comparing model performance across multiple tasks: referring segmentation, disaster scene recognition, damaged road area estimation, and damaged object relational reasoning.

For the referring segmentation task with the prompt "Identify and segment the roads with debris blockage and segment their regions," the optical-SAR modality combination proves more challenging than traditional optical-optical segmentation due to the inherent differences in sensor characteristics. All baseline methods fail to accurately identify and segment the debris-affected road regions. Notably, even fine-tuned LISA produces no viable segmentation outputs for this complex cross-modal scenario. Although fine-tuned PSALM demonstrates partial success by correctly segmenting one debris-blocked road section, significant performance gaps remain that warrant further investigation.

The disaster scene recognition and damaged road area estimation tasks exhibit performance trends consistent with those demonstrated in Fig. 12, where baseline VLMs show limited capability while fine-tuned models achieve substantially better results.

For the damaged object relational reasoning task with the prompt "Explain how the object in the red box spatially relates to the object in the yellow box," the challenge intensifies considerably when working with SAR imagery. This increased difficulty stems from the substantial domain gap between SAR and optical data, as well as the reduced spectral information available in SAR images for object identification and spatial reasoning. Among all evaluated models, only the fine-tuned InternVL3-8B successfully provides accurate spatial relationship descriptions.

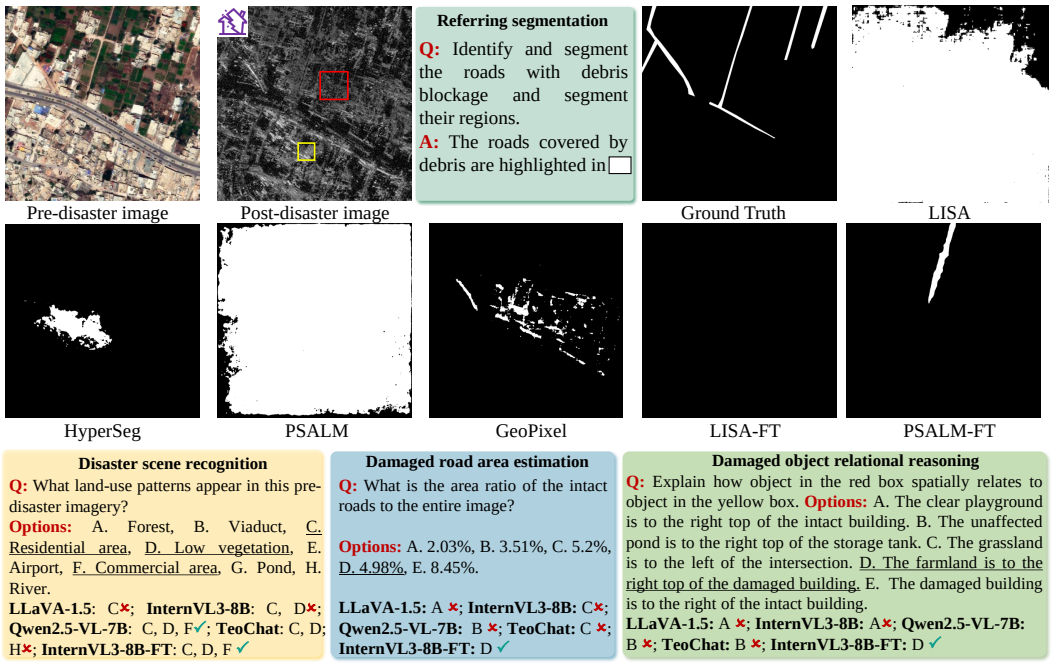

Figure 13: Visualization of compared predicted results for earthquake disaster scene under the optical-SAR setting.

I Broader impacts

The DisasterM3 dataset has significant potential for positive societal impact by enhancing disaster response capabilities through more accurate and timely damage assessment. By enabling vision-language models to better understand disaster scenarios, this work could help emergency responders prioritize affected areas, allocate resources more efficiently, and accelerate recovery planning, potentially saving lives and reducing economic losses. The multi-sensor approach is particularly valuable for developing comprehensive situational awareness during extreme weather events when optical sensors are compromised. However, there's also the risk of over-reliance on AI systems during critical emergency situations, where incorrect assessments could lead to misallocation of vital resources. To mitigate this concern, we recommend that DisasterM3-trained models be deployed as assistive tools alongside human experts rather than autonomous decision-makers in emergency management scenarios.