次世代災害対応のための視覚言語モデルの構築

Naoto Yokoya, Junjue Wang, Weihao Xuan
Graduate School of Frontier Sciences, The University of Tokyo

UTokyo

Background

Human efforts

1. Damage assessment and annotation (building, road, factory, etc).
2. Rescue route planning.
3. Emergency shelter area arrangement.
4. Allocation of personnel and supplies according to rescue priorities.
5. Restoration advices.

...

Motivation: **How to leverage large vision-language models for artificial intelligent disaster assessment and response?**

Research Plan

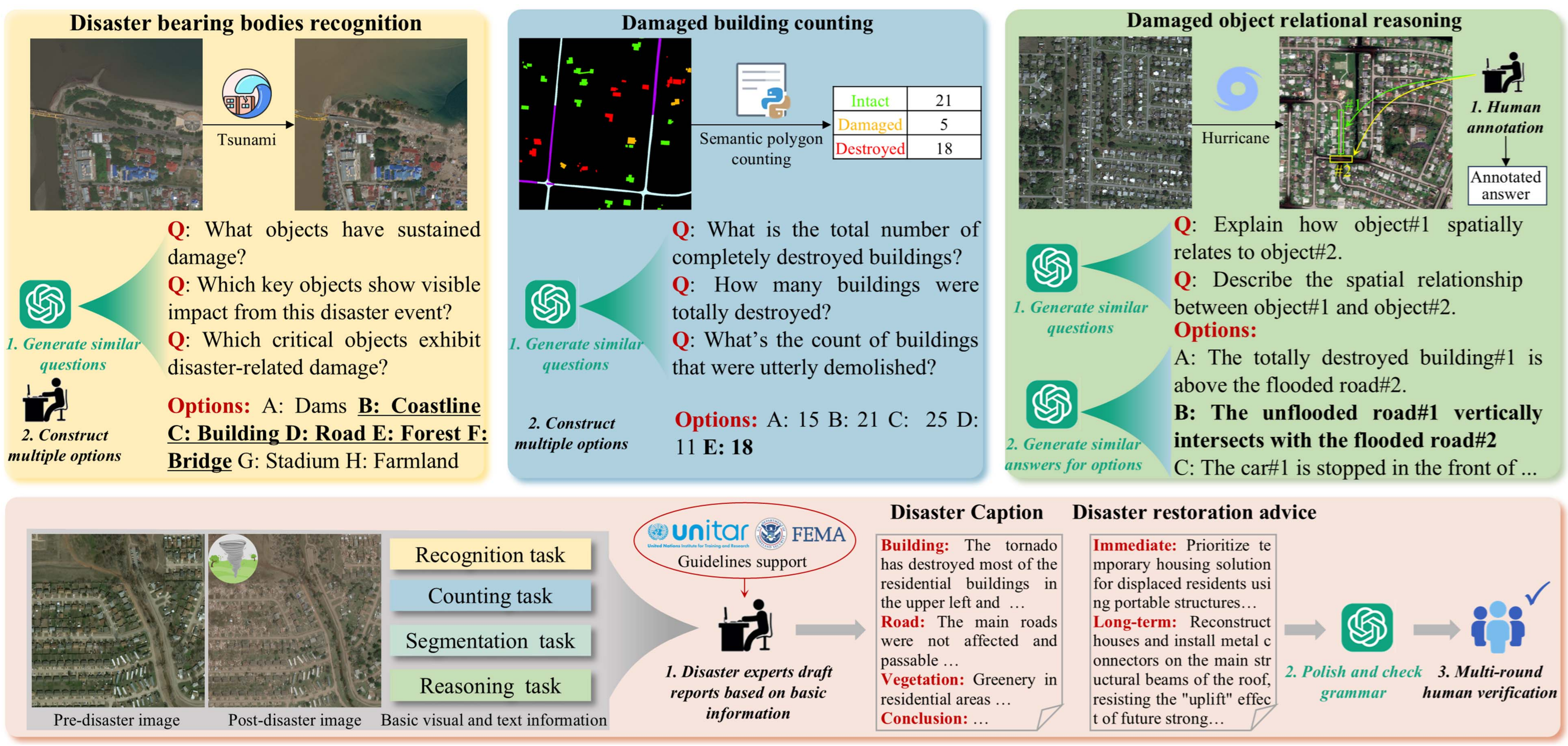
Goal1: Curate a large-scale vision-language dataset for disaster damage assessment and response.

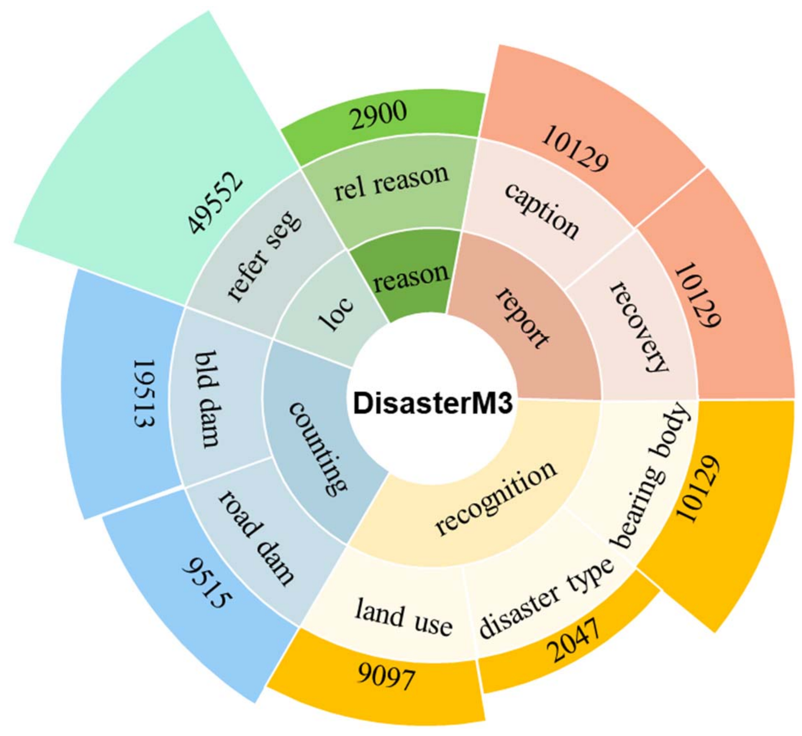
Step1.1： Construct a semi-automatic pipeline for high-quality annotation. → ***Done***

Step1.2： Benchmark open-source and commercial vision-language models disaster response. → ***Ongoing***

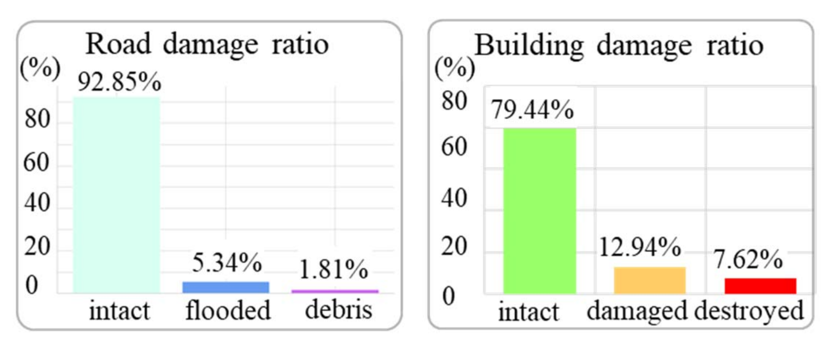
Goal2: Develop a disaster assistant for conversation-based disaster response. → ***Future plan***

DisasterM3: Multi-hazard, multi-sensor, multi-task vision language dataset





(a) Sample distributions



(b) Damage-level distributions

Preliminary Experiments

Method	Accuracy (%)							Disaster Caption				Restoration Advice			
	AVG	LUC	DTR	BBD	BDC	DRE	ORR	AVG	DAP	DDR	FC	AVG	RNR	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 [53]	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 [53]	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 [53]	39.3	43.5	72.5	5.3	29.4	28.7	<u>56.1</u>	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															
GPT4o [12]	39.3	<u>49.4</u>	<u>80.5</u>	10.6	24.2	21.4	49.8	2.27	2.25	2.28	2.28	<u>3.19</u>	<u>3.92</u>	<u>2.95</u>	2.69
GPT4.1 [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**	<u>21.5</u>	<u>34.3</u>	<u>29.4</u>	36.2	**3.90**	**3.76**	**3.53**	**4.41**	3.11	3.73	2.88	<u>2.73</u>
Δ	↑9.2	↑9.4	↑17.0	↑16.8	↑0.1	↑0.1	↑12.3	↑2.15	↑2.07	↑1.82	↑2.56	↑1.26	↑1.20	↑1.83	↑1.24
InternVL3-8B [53]	<u>41.7</u>	42.6	79.3	**23.9**	29.1	24.9	50.6	<u>3.83</u>	<u>3.69</u>	<u>3.49</u>	<u>4.32</u>	**3.31**	3.92	**3.10**	**2.90**
Δ	↑10.4	↑3.0	↑25.8	↑19.9	↓-1.2	↑0.8	↑14.4	↑1.87	↑1.81	↑1.57	↑2.23	↑0.56	↑0.40	↑0.57	↑0.69

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), and object relational reasoning (ORR). The open-ended tasks are scored using GPT4.1 on 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.

Domain gap for disaster scenarios. Existing VLMs show significant domain gaps when processing disaster scenes, limiting their performance.

Larger VLMs achieve higher performances. Scaling laws hold for VLMs: larger models perform better. Commercial models excel due to massive training data.

Fine-tuned models improve comprehensively. DisasterM3 fine-tuning significantly improves VLM performance and narrows domain gaps. Disaster-specific terminology enhances report quality.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, Naoto Yokoya. DisasterM3: A Remote Sensing Vision-Language Dataset for Disaster Damage Assessment and Response[J]. arXiv preprint arXiv:2505.21089, 2025.