Multimodal Large Language Model Meets Remote Sensing

GISLab Short-Term Course 2025 Summer

Zhenyuan Chen

School of Earth Science, Zhejiang University

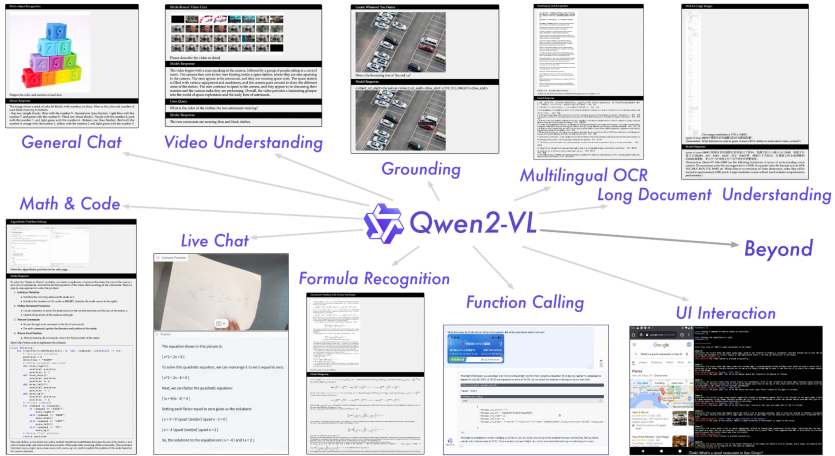
2025

bili_sakura@zju.edu.cn

Outline

1. An Introduction to Multimodal Large Language Models (MLLMs)
2. Remote Sensing Change Captioning
3. Project: Bi-Temporal Captioning with RSCC

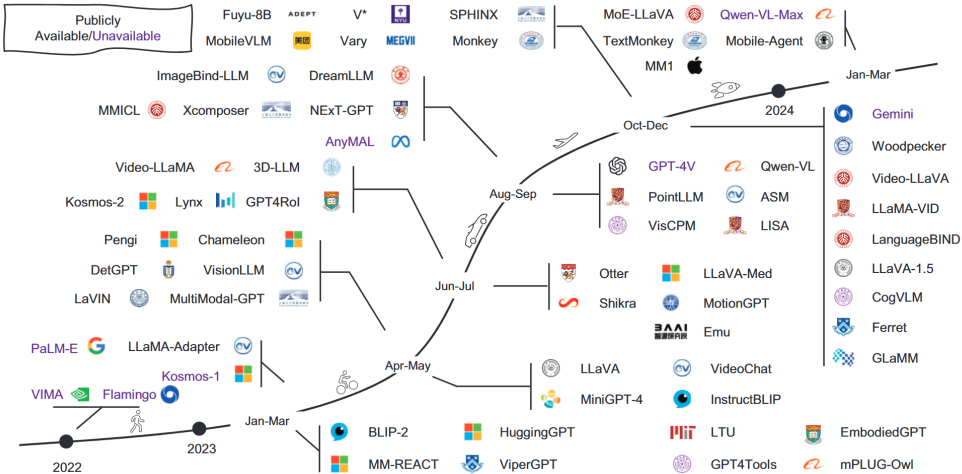
Background: Multimodal Large Language Models (MLLMs)



Qwen2-VL (Wang et al., 2024).

Wang, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, arXiv, 2024.

Background



MLLM Timeline (Yin et al., 2024).

Yin, et al. A Survey on Multimodal Large Language Models. National Science Review. 2024.

What is a Multimodal Large Language Model (MLLM)?

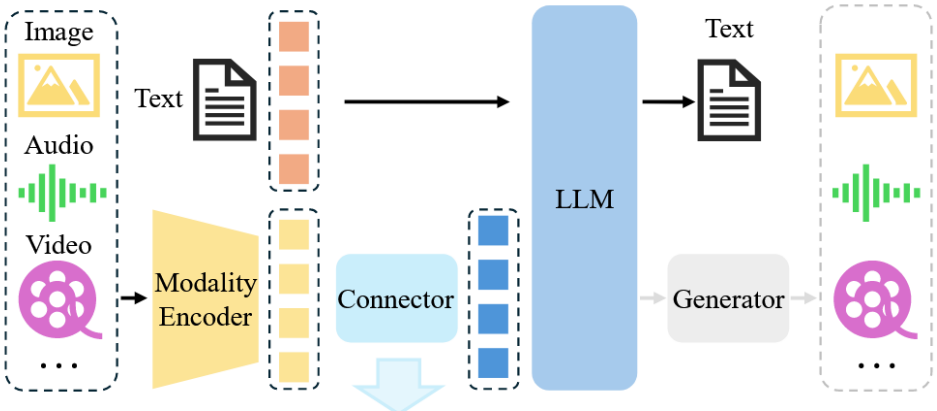


Figure: General architecture of a Multimodal Large Language Model (MLLM) (Yin et al., 2024).

Yin, et al. A Survey on Multimodal Large Language Models. National Science Review. 2024.

Vision Encoder: CLIP

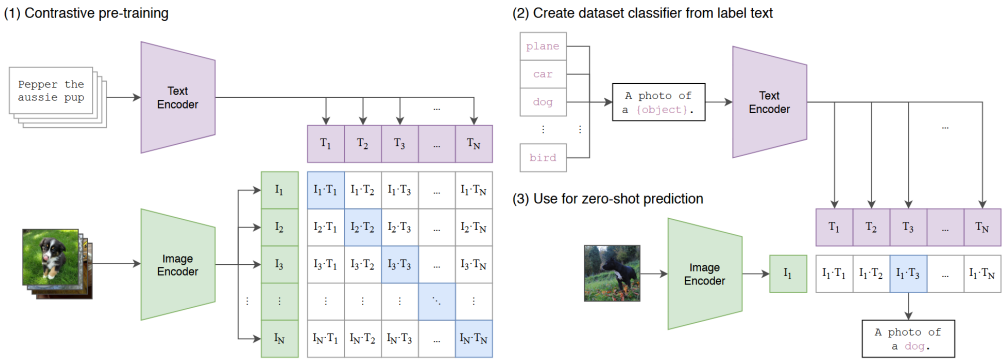
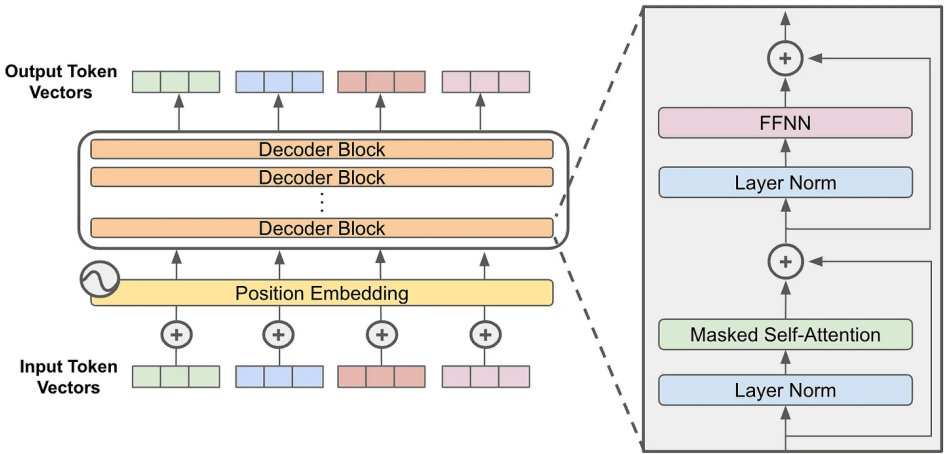


Figure: OpenAI CLIP architecture (Radford et al., 2021).

Radford, et al. Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021.

LLM Backbone: Large Language Models

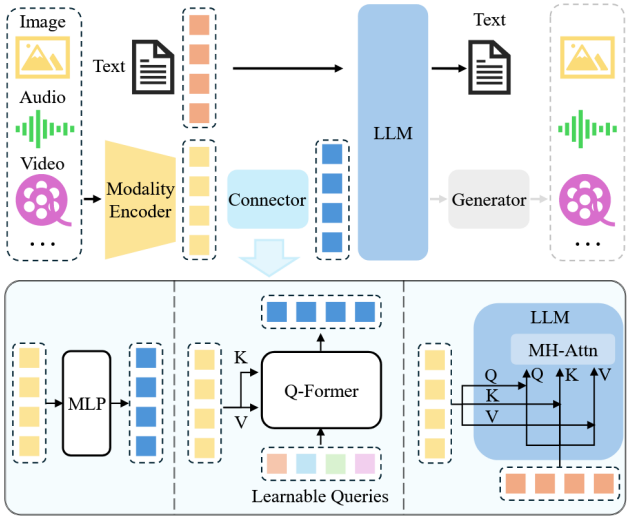


Latest LLMs are generally composed of a series of Transformers decoder blocks (Vaswani et al., 2017).

(image source:cameronrwolfe)

Vaswani, et al. Attention Is All You Need, NeurIPS, 2017.

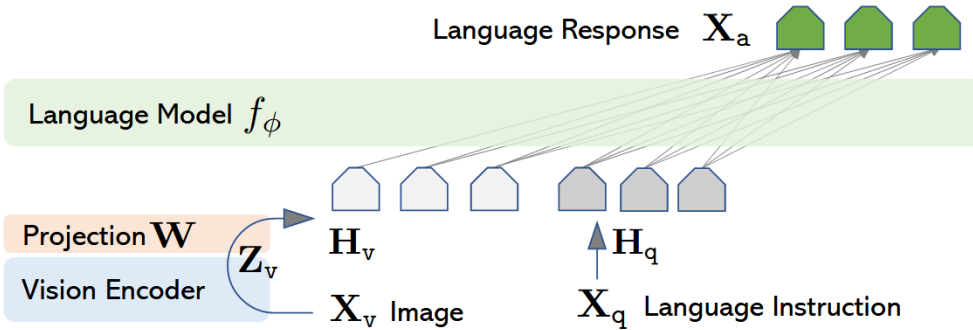
Connector: General MLLM



An typical MLLM architecture includes an encoder, a connector, and a LLM. (Yin et al., 2024).

Yin, et al. A Survey on Multimodal Large Language Models. National Science Review. 2024.

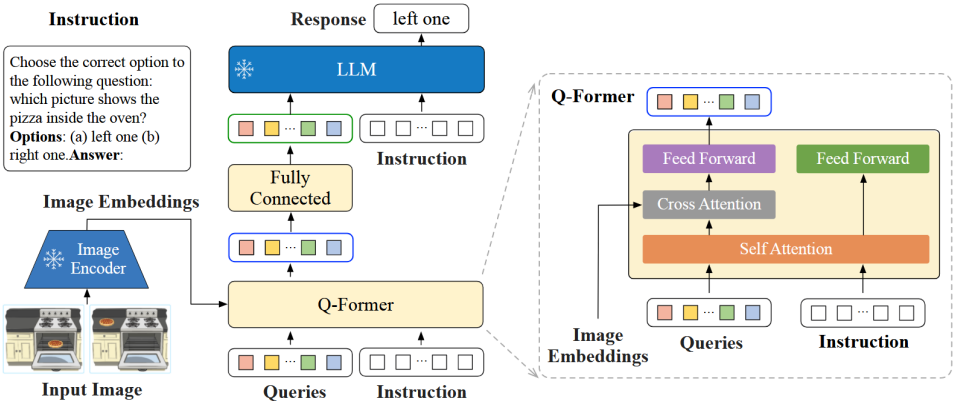
Connector: MLP



Connector architecture in LLaVA (Liu et al., 2023).

Liu, et al. Visual Instruction Tuning, NeurIPS, 2023.

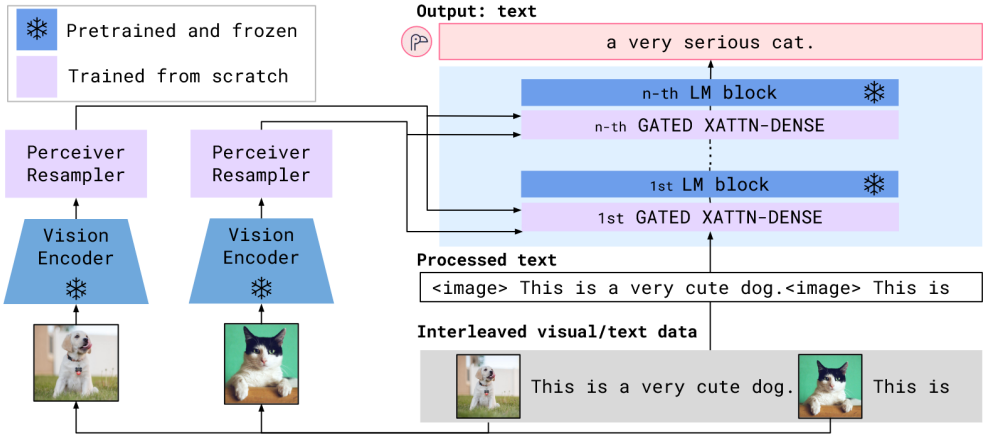
Connector: Q-Former



Connector architecture in InstructBLIP (Dai et al., 2023).

Dai, et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, NeurIPS, 2023.

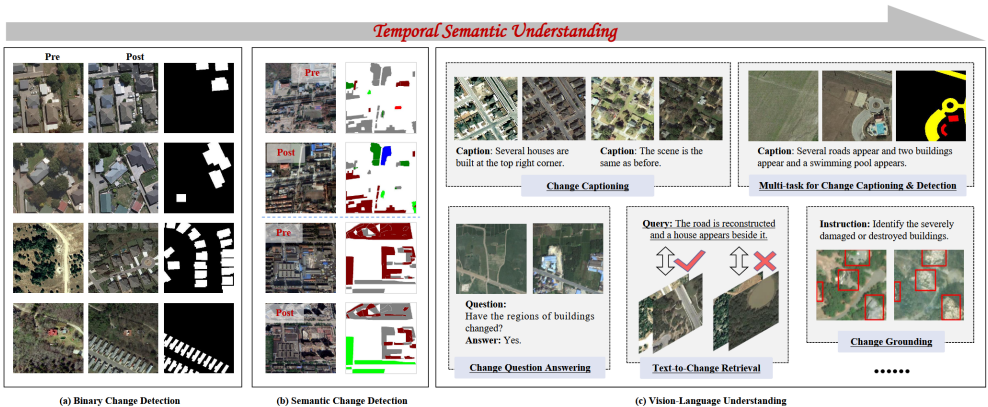
Connector: Feature-level Fusion



Connector architecture in Flamingo (Alayrac et al., 2022).

Alayrac, et al. Flamingo: A Visual Language Model for Few-Shot Learning, NeurIPS, 2022.

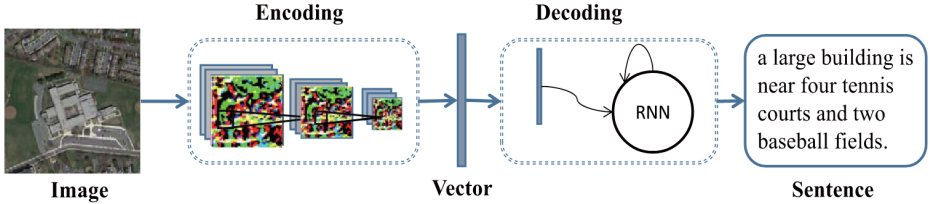
Remote Sensing Temporal Tasks



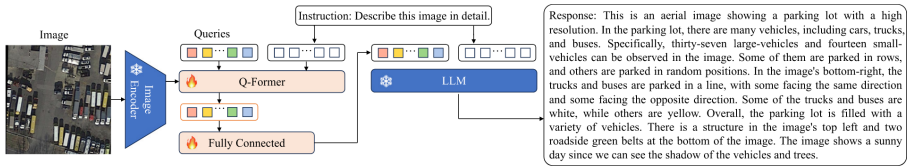
Three concepts in temporal image understanding. (Liu et al., 2024)

Liu, et al. Remote Sensing Temporal Vision-Language Models: A Comprehensive Survey, arXiv, 2024.

Remote Sensing Captioning



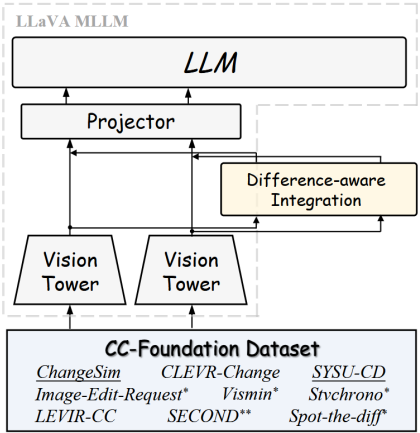
Outline of encoder–decoder for remote sensing image caption. (Lu et al., 2018)



RS-GPT. (Hu et al., 2025)

Lu, et al. Exploring Models and Data for Remote Sensing Image Caption Generation. TGRS. 2018.
Hu, et al. RSGPT: A remote sensing vision language model and benchmark. ISPRS. 2025.

Remote Sensing Change Captioning

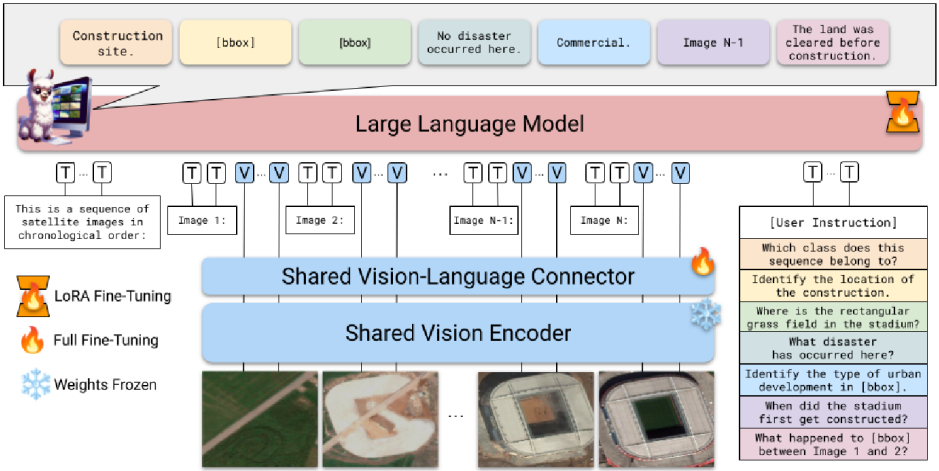


GT: Two rows of houses are built on both sides of the road and trees disappear.
MCCFormer: *Many houses are built along the road.*
RSICCFormer: *Massive houses along the roads appear in the desert.*
CCExpert: *Two rows of houses are built on both sides of the road.*

CCExpert framework and results. (Wang et al., 2024)

Wang, et al. CCExpert: Advancing MLLM Capability in Remote Sensing Change Captioning with Difference-Aware Integration and a Foundational Dataset, arXiv, 2024.

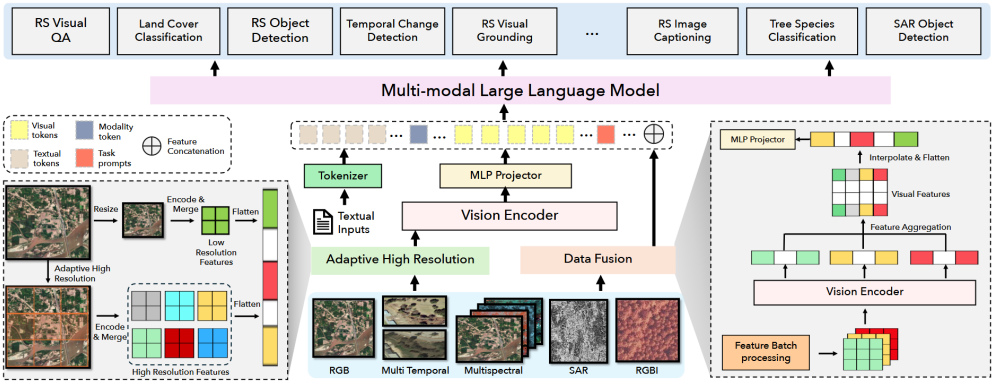
Remote Sensing Foundation Models



TEOChat. (Irvin et al., 2025)

Irvin, et al. TEOChat: A Large Vision-Language Assistant for Temporal Earth Observation Data, ICLR, 2025.

Remote Sensing Foundation Models



EarthDial. (Soni et al., 2025)

Soni, et al. EarthDial: Turning Multi-sensory Earth Observations to Interactive Dialogues, CVPR, 2025.

Project

Dataset

Change Caption

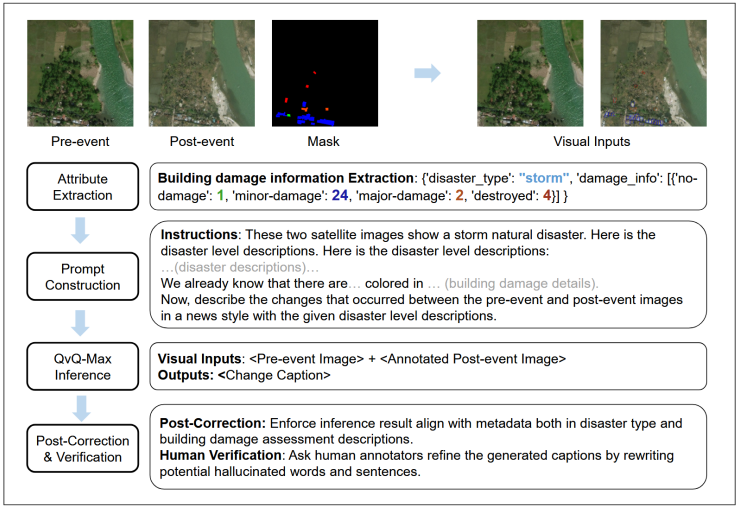
Event: Flooding

The area experienced a significant flood transformation, as evidenced by the pre-event image showing clear, undisturbed land with a single intact building, while the post-event image reveals the same location now submerged under murky floodwaters, with the building surrounded by water, indicating a shift from Disaster Level 0 to Disaster Level 2 conditions, highlighting the severe impact of the natural disaster on the infrastructure.

Example of RSCC (Chen et al., 2025).

Chen, et al. RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events, 2025.

Pipeline



Pipeline of constructing RSCC dataset (Chen et al., 2025).

Chen, et al. RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events, 2025.

Benchmark

Table 3: Detailed image caption performance on the subset of RSCC dataset (naive/zero-shot results). Avg_L denotes the average word number of generated captions. **Boldface** indicates the best performance while underline denotes the suboptimal performance. *BLIP-3 and LLaVA-OneVision tend to repeat their answer endlessly, which cause large caption lengths.

Model	N-Gram		Contextual Similarity	Avg_L
(#activate params)	ROUGE(%)↑	METEOR(%)↑	ST5-SCS(%)↑	
BLIP-3 (3B) [74]	4.53	10.85	44.05	*456
Kimi-VL (3B)[26]	12.47	16.95	51.35	87
Phi-4-Multimodal (4B) [42]	4.09	1.45	34.55	7
Qwen2-VL (7B)[69]	11.02	9.95	45.55	42
LLaVA-NeXT-Interleave (8B) [30]	12.51	13.29	46.99	57
LLaVA-OneVision (8B)[29]	8.40	10.97	46.15	*221
InternVL 3 (8B) [78]	12.76	15.77	51.84	64
Pixtral (12B) [43]	12.34	15.94	49.36	70
CCExpert (7B) [71]	7.61	4.32	40.81	12
TEOChat (7B)[24]	7.86	5.77	52.64	15
Ours (7B)	**14.99**	**16.05**	**58.52**	44

RSCC benchmark (Chen et al., 2025).

Chen, et al. RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events, 2025.

Result

(a)

Kimi-VL (3B): The two satellite images depict a river flowing through a dense forest. In the first image, the river is bordered by a road on one side, with a few buildings and structures visible along the riverbank. The second image shows a similar scene, but the road and buildings are not visible, providing a clearer view of the forest and the river. The overall landscape in both images is characterized by lush greenery and a tranquil natural setting.

GT: In a recent comparison of satellite imagery, significant changes have been observed, the water level has risen notably, submerging parts of the land previously visible, altering the landscape's appearance dramatically.

Qualitative results. (Chen et al., 2025).

Chen, et al. RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events, 2025.

Result

(b)

BLIP-3 (4B): The first satellite image shows a large body of water with a few small islands scattered around it. The second image shows the same area, but with a much larger and more detailed view of the islands and the surrounding water. The second image also shows more detail in the vegetation and terrain of the islands. (Repeat.....)

GT: A dramatic transformation has occurred, the once lush green golf course now appears scorched and barren, the vibrant grass replaced by patches of dark earth, the clear pathways now marred by signs of destruction.

Qualitative results. (Chen et al., 2025).

Chen, et al. RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events, 2025.

Result

(c)

Phi-4-Multimodal (4B): The second image has more buildings.
Qwen-2-VL (7B): The image shows a significant change in the area, with many houses destroyed or damaged, and debris scattered across the streets.
TEOChat (7B): A hurricane has struck the area, resulting in significant damage.
CCExpert (7B): The area has been cleared of several buildings.

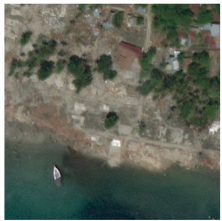
GT: A devastating transformation has occurred, as the once lush, green neighborhood is now replaced by a barren landscape of debris and destruction, highlighting the severe impact of a catastrophic event, with every structure reduced to rubble, and the streets eerily empty of life.

Qualitative results. (Chen et al., 2025).

Chen, et al. RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events, 2025.

Result

(d)



LLaVA-Interleave (7B): The first image shows a clear view of a coastal town with houses, roads, and green areas. The second image has been altered to show the town with a significant portion of the landscape covered in debris, suggesting a recent disaster or extensive destruction. The boat that was previously visible in the water is no longer seen in the second image.

GT: In a striking transformation, the once densely populated coastal area now appears significantly cleared, revealing patches of bare earth and fewer structures, while the waterway shows a reduction in visible boats, indicating a notable shift in human activity and environmental conditions.

Qualitative results. (Chen et al., 2025).

Chen, et al. RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events, 2025.