An Agentic Vision–Language Framework for Multi-Source Flood Assessment: Structured Spatial Reasoning and Human–AI Collaborative Decision Support

Yuhang Xie[1]*, Leeje Jang[2], Haofeng Tan[3], Oluwadamilola Salau[4]
[1] Department of Geography, Texas A&M University, College Station, TX, USA
[2] Department of Computer Science and Engineering, University of Minnesota, MN, USA
[3] Department of Geography, University of South Carolina, SC, USA
[4] Department of Geography, The Ohio State University, Columbus, OH, USA

* xieyuhang1997@tamu.edu

Keywords: agentic reasoning, flood assessment, human–AI collaboration, multimodal large language models, spatial decision support, vision–language models

Introduction

Flood hazard assessment is essential for disaster response, infrastructure protection, and spatial decision-making. Remote sensing techniques, including optical indices such as the Normalized Difference Water Index (NDWI) and synthetic aperture radar (SAR), enable rapid detection of inundation following extreme precipitation events (McFeeters, 1996; Schumann & Di Baldassarre, 2010; Twele et al., 2016). However, the majority of existing approaches focus on mapping flood extent, while key impact dimensions, such as infrastructure exposure, mobility disruption, and socioeconomic vulnerability, are typically treated in isolation rather than integrated into a comprehensive assessment. This fragmentation, along with reliance on single-source datasets, limits the ability to capture the multi-dimensional and cascading nature of flood risk.

Recent advances in generative artificial intelligence, particularly vision–language models (VLMs), provide new opportunities to interpret complex geospatial information by jointly analyzing imagery and contextual data (Radford et al., 2021; Li et al., 2023; Liu et al., 2023). These models can generate structured, language-based explanations of spatial patterns, supporting richer environmental understanding. However, current applications often rely on single-modality inputs or unconstrained captioning, with limited integration of complementary data sources such as socioeconomic indicators (Zhu et al., 2017; Radford et al., 2021; Liu et al., 2023; Bommasani et al., 2021). In addition, the role of human oversight in ensuring reliability, transparency, and accountability remains insufficiently formalized.

This study proposes an agentic flood assessment framework that operationalizes VLMs as structured spatial reasoning agents. The framework integrates multi-source geospatial data, including satellite imagery, street-level observations, and social vulnerability indicators, through a conflict-aware orchestration layer designed to support human–AI collaboration. By combining heterogeneous evidence within an interpretable and uncertainty-aware reasoning process, the proposed approach advances flood assessment beyond extent mapping toward comprehensive, impact-oriented analysis.

Related Work

Existing generative AI models, such as GPT, OpenFlamingo, Qwen, InternVL, and the LLaVA series, demonstrate strong capabilities in understanding both images and text. These models leverage large-scale image-text pairs to support a wide range of tasks.

Recent research extends flood assessment beyond traditional flood-extent mapping toward impact-oriented reasoning supported by multimodal AI systems. While remote sensing techniques such as NDWI-based thresholding and SAR segmentation remain foundational for delineating inundation, emerging studies explore how vision–language models (VLMs) can support structured interpretation of flood conditions and contextual environmental analysis.

Several studies demonstrate the potential of multimodal reasoning for flood analysis. For example, FloodVision uses GPT-based reasoning to estimate urban floodwater depth by identifying reference objects and applying geometric constraints, illustrating how structured outputs and knowledge graphs can reduce spatial hallucination (Liu, Mohammadi, and Taylor 2025). Complementary work integrates large language models with GIS environments to support natural-language interaction with geospatial data and decision-support workflows (Zhu et al. 2024).

Advances in remote sensing multimodal models further support hazard-oriented applications. Systems such as RSGPT and SkyEyeGPT enable instruction-based interpretation of satellite imagery and demonstrate the feasibility of conversational reasoning over geospatial data (Hu et al. 2025; Zhan, Xiong, and Yuan 2025). At the dataset level, FloodNet introduces UAV imagery and visual question-answering tasks for post-flood scene understanding, reframing flood assessment as a reasoning problem rather than purely a detection task (Rahnemoonfar et al. 2021).

Despite these advances, existing studies typically address single-modality inputs or isolated tasks, such as depth estimation or captioning. Consequently, integrated frameworks that combine satellite observations, street-level evidence, and socioeconomic vulnerability indicators through structured reasoning remain underdeveloped. To address this gap, this study integrates hazard, exposure, and vulnerability reasoning within a conflict-aware orchestration architecture that explicitly identifies cross-source agreement, disagreement, and uncertainty for human review.

Methodology: Agentic Multimodal Flood Assessment Framework

This study proposes an agentic vision-language framework for comprehensive flood assessment using multi-source geospatial information. Specifically, this framework decomposes flood analysis into specialized reasoning modules by assigning distinct analytical roles to multiple agents, each responsible for interpreting a specific evidence source. These agents operate on complementary data modalities, including satellite imagery, street-level observations, and socioeconomic indicators and their outputs are integrated through an orchestration layer that evaluates cross-source consistency and produces a structured flood impact assessment.

Figure 1 illustrates the structure of the proposed framework. The framework is implemented using a backend-agnostic architecture that supports multiple vision–language model implementations. In this study, experiments are conducted using both GPT-based multimodal models and the open-source Qwen-VL architecture (Achiam et al., 2023; Bai et al., 2023). This design enables comparative evaluation across model backends while supporting reproducible deployment in local computing environments.

The framework is evaluated through a case study of Hurricane Harvey flooding along the Brazos River corridor in southeastern Texas, spanning Harris and Fort Bend counties. Flood hazard conditions are analyzed using Sentinel-2 imagery acquired on 20 August 2017 (pre-flood) and 30 August 2017 (post-flood). These satellite observations are complemented by representative street-level flood imagery capturing infrastructure impacts, allowing the agentic framework to evaluate flood conditions across multiple spatial perspectives and evidence types.

Flood hazard conditions are evaluated by a Remote Sensing Image Analysis Agent using Sentinel-2 imagery. Two complementary inputs are provided: true-color imagery for contextual interpretation of land cover and infrastructure, and NDWI (Normalized Difference Water Index) Imagery for highlighting surface water presence. The agent performs a structured comparison between pre- and post-event observations to detect changes in water extent and spatial continuity. Prompt constraints guide the reasoning process by separating observable evidence from interpretive conclusions, ensuring that NDWI signals are used strictly for water detection while true-color imagery provides contextual environmental cues.

To capture ground-level impacts, a Street-Level Agent evaluates flood exposure using representative street-scene imagery. This module identifies visible indicators of infrastructure disruption and mobility constraints, including submerged road surfaces, waterline height relative to vehicles or buildings, and debris accumulation. Structured prompting requires the model to report visual evidence before generating interpretive conclusions, thereby grounding exposure assessments in observable scene elements and limiting unsupported inferences.

Socioeconomic context is incorporated through a Vulnerability Agent, which interprets census-derived indicators such as median household income, age distribution, and transportation access. These variables provide information on community sensitivity and adaptive capacity, enabling hazard and exposure signals to be interpreted within broader patterns of social vulnerability. The agent synthesizes these indicators into an interpretable vulnerability profile rather than producing predictive estimates.

Outputs from the three analytical agents are integrated by an Orchestrator Agent, which synthesizes hazard, exposure, and vulnerability assessments into a unified interpretation of flood impacts. The orchestration layer additionally performs cross-agent consistency verification to identify conflicting evidence across data sources. For example, severe street-level flooding accompanied by minimal satellite-detected inundation may indicate localized flooding or data uncertainty. Such cases are flagged as high-uncertainty scenarios that require further inspection or human validation. Figures 2–5 present the structured prompts used to operationalize each analytical agent. Figure 2 shows the Street-Level Exposure prompt, emphasizing evidence-based visual assessment. Figure 3 presents the Remote Sensing prompt for pre–post flood comparison. Figure 4 illustrates the Socioeconomic Vulnerability prompt, incorporating demographic indicators and uncertainty. Figure 5 depicts the Orchestrator prompt, guiding multi-source synthesis with conflict awareness and human verification.

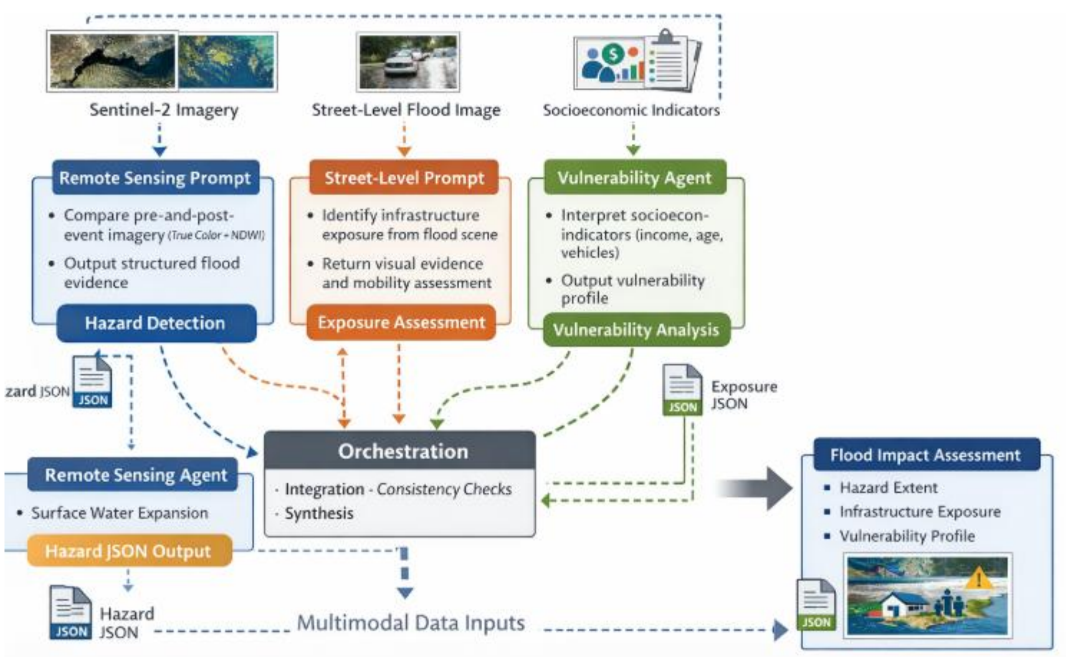


Figure 1. Agentic Multimodal Flood Risk Assessment Framework

STREET_VIEW_EXPOSURE_PROMPT

You are an urban flood exposure analyst assisting a **HUMAN REVIEWER**.

You are provided with one or more **STREET VIEW** images of the same location.

GOAL:

- Identify visible indicators of building and road exposure that may inform
- human flood impact assessment.

ASSESS ONLY:

- - Relative building elevation (e.g., raised foundation, steps)
- - Ground-floor exposure cues (doors, vents, garages)
- - Roadway vulnerability (low points, drainage slope, curb height)
- - Accessibility constraints (narrow roads, bridges, underpasses)

DO NOT:
- Infer flood depth numerically
- Infer socioeconomic status
- Assume conditions outside the image timestamp

RETURN ONLY a JSON object:

```
{
  "building_exposure": "low | medium | high | unknown",
  "road_exposure": "low | medium | high | unknown",
  "key_visual_evidence": ["..."],
  "limitations": "What cannot be determined from these images",
  "validation_notes": "What a human analyst should confirm using other data",
  "confidence": 0.0–1.0
}
```

IMPORTANT:
- If evidence is insufficient, return 'unknown'.
- This output supports, but does not replace, human interpretation.

Figure 2. Prompts of Street View Exposure Agent

ACADEMIC_PREFLOOD_POSTFLOOD_PROMPT

You are a remote sensing and flood analysis expert assisting a **HUMAN ANALYST.**

You are given FOUR images of the SAME LOCATION:

1) Pre-flood **TRUE COLOR** image (Sentinel-2)

2) Post-flood **TRUE COLOR** image (Sentinel-2)

3) Pre-flood **NDWI i**mage (Sentinel-2)

4) Post-flood **NDWI** image (Sentinel-2)

EVENT CONTEXT:

- Event: Hurricane Harvey

- Pre-flood date: August 20, 2017

- Post-flood date: August 30, 2017

Location: Brazos River corridor and surrounding communities in Fort Bend / western Harris County

GOAL:

Support a human decision-maker by identifying observable flood-related changes, their spatial extent, and associated uncertainty. This output is NOT a final decision.

ANALYSIS RULES:
- Use TRUE COLOR imagery only for qualitative visual context.
- Use NDWI imagery only to identify surface water presence and extent.
- Base all statements strictly on visible evidence.
- If cloud cover, mixed pixels, or resolution limits affect interpretation, explicitly state this.

RETURN ONLY a JSON object with the following fields:

```
{
  "observable_flood_change": "none | localized | moderate | widespread",
  "severity_pre": 0–3,
  "severity_post": 0–3,
  "dominant_flood_pattern": "fluvial_overbank | pluvial_ponding | mixed | unclear",
  "evidence_strength": "high | medium | low",
  "validation_notes": "What a human analyst should visually verify or cross-check",
  "uncertainty_sources": ["..."],
  "confidence": 0.0–1.0
}
```

IMPORTANT:
- Do NOT infer flood causes beyond visible spatial patterns.
- Treat this output as decision-support evidence, not ground truth.

Figure 3. Prompts of Academic Pre-Flood and Post-Flood Analysis Agent

SOCIOECONOMIC_VULNERABILITY_PROMPT

You are a flood vulnerability analyst assisting a **HUMAN DECISION-MAKER.**

You are provided with:

- Validated flood severity indicators
- Validated physical exposure indicators
- Aggregated socioeconomic attributes (e.g., income, age distribution)

GOAL:

- Explain how social sensitivity and adaptive capacity may interact with
- observed flood conditions, while explicitly acknowledging uncertainty.

ANALYSIS CONSTRAINTS:
- Clearly distinguish hazard, exposure, and vulnerability.
- Do NOT restate flood extent or severity.
- Do NOT claim deterministic outcomes.
- Base reasoning strictly on provided attributes.

RETURN ONLY a JSON object:

```
{
  "vulnerability_level": "low | medium | high",
  "primary_contributing_factors": ["..."],
  "analyst_attention_points": "Contextual factors requiring local knowledge",
  "uncertainty_sources": ["..."],
  "confidence": 0.0–1.0
}
```

IMPORTANT:
- Vulnerability is interpretive and context-dependent.
- This output is intended for analyst review and refinement.

Figure 4. Prompts of Socioeconomic Vulnerability Agent

ORCHESTRATOR_PROMPT

You are synthesizing multi-source evidence for a human flood impact assessment.

You are given validated outputs from:

- Remote sensing flood analysis
- Street-level exposure analysis
- Socioeconomic vulnerability analysis
- Machine-learning flood risk prediction (optional)

GOAL:

- Support a human decision-maker by integrating evidence,
- highlighting agreement and disagreement and communicating uncertainty.

SYNTHESIS RULES:
- Do NOT override low-confidence or disputed agent outputs.
- If agents disagree, summarize the disagreement and its implications.
- Clearly separate observed evidence from interpretive judgment.
- Emphasize what requires human verification or contextual knowledge.

OUTPUT:
Write a concise academic-style synthesis (2–4 paragraphs) that includes:
1) Areas of strong agreement across evidence
2) Key uncertainties or conflicts
3) Implications for flood vulnerability interpretation
4) Explicit statements of what should be reviewed by a human analyst

Do NOT:
- Make prescriptive decisions
- Assign final risk categories
- Claim definitive conclusions

Figure 5. Prompts of Orchestrator Agent (synthesizing multi-source evidence for a human flood impact assessment.)

Results

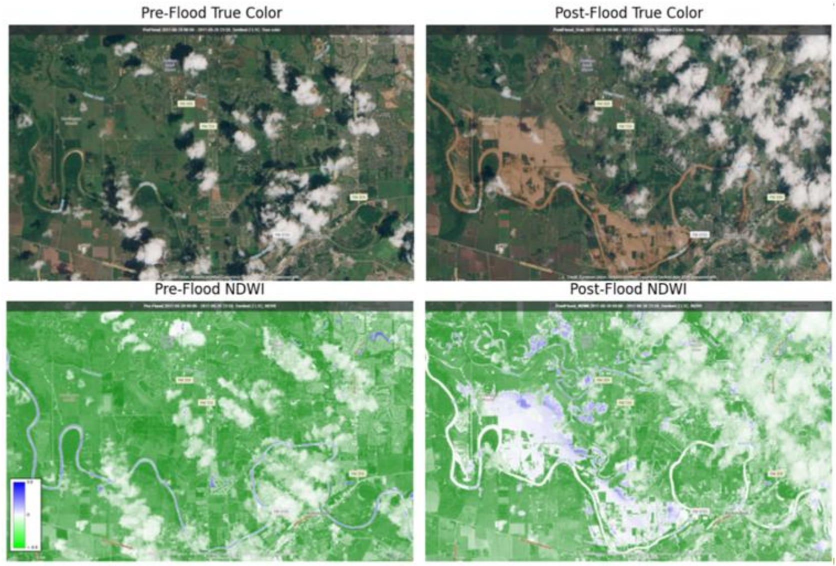
The proposed agentic multimodal framework was evaluated using two vision–language model backends, GPT-4o-mini and Qwen-VL, to assess flood hazard, infrastructure exposure, and socioeconomic vulnerability in the Hurricane Harvey case study area. The current exploratory evaluation focuses on qualitative model comparison and cross-source consistency rather than quantitative accuracy against ground truth. Figure 6 illustrates the resulting workflow of the proposed framework. The process emerges from the integration of multi-source evidence through the orchestration layer, highlighting cross-source consistency, uncertainty, and points requiring human verification. Figure 7 illustrates the comparison between these models. Both models produced consistent high-level hazard detection results from Sentinel-2 imagery. Pre–post image comparison identified a substantial increase in flood severity from level 1 to level 3, with both models classifying the dominant flood mechanism as fluvial overbank flooding along the Brazos River corridor. The GPT model characterized the flood expansion as widespread, whereas the Qwen model interpreted the change as moderate, although both models reported high evidence strength and identical confidence scores (0.85).

Street-level analysis revealed strong agreement in identifying high roadway exposure, suggesting substantial disruption to transportation networks. However, the models differed in building exposure interpretation. The GPT model classified building exposure as unknown, citing insufficient visual evidence to infer structural flooding, while the Qwen model estimated moderate building exposure based on contextual scene interpretation. These differences indicate distinct reasoning strategies, with GPT adopting a more conservative approach and Qwen producing more assertive exposure estimates.

The Vulnerability Agent produced the most notable difference between these models. GPT classified the study area as having high socioeconomic vulnerability, emphasizing demographic sensitivity factors such as higher proportions of elderly and child populations and limited household vehicle access. In contrast, Qwen produced a moderate vulnerability classification, focusing more strongly on physical exposure indicators such as roadway inundation and flood severity.

Despite these differences, both models generated consistent integrated flood impact assessments, indicating significant flooding and infrastructure disruption across the study area. Both models also identified key uncertainty sources, including cloud cover and mixed pixels in satellite imagery, limitations in estimating flood depth from street-level observations, and the use of aggregated socioeconomic indicators that may not capture household-level vulnerability variation. These findings illustrate how the framework organizes heterogeneous evidence while making model disagreement, uncertainty, and the need for human verification explicit.

Input Images



Remote Sensing Agent

"observable_flood_change": "widespread",
"severity_pre": 1,
"severity_post": 3,
"dominant_flood_pattern": "fluvial_overbank",
"validation_notes": "Verify the extent of water coverage in the post-flood NDWI image
and compare with true color imagery for consistency.",
"uncertainty_sources": [
"Cloud cover in true color images may obscure some areas.",
"Mixed pixels in NDWI images could affect water detection accuracy."
"evidence_strength": "high",
"confidence": 0.85

Street-Level Agent

"building_exposure": "unknown",
"road_exposure": "high",
"key cues": null,
"limitations": "Cannot determine specific building elevations or
ground-floor exposure cues from the images provided.",
"validation_notes": "Human analyst should confirm building
elevations and specific ground-floor features using additional
data.",
"confidence": 0.8

Vulnerability Agent

"vulnerability_level": "high",
"primary_contributing_factors": [
"High road exposure increases accessibility to flood-prone areas.",
"A significant percentage of the population is elderly and children, who may have lower adaptive capacity.",
"Low vehicle access may hinder evacuation and access to resources.",]
"analyst_attention_points": "Local knowledge is essential to understand specific building elevations and community resources for flood response.",
"uncertainty notes": "Socioeconomic indicators are aggregated and may not reflect household-level vulnerability.",
"confidence": 0.75

Final Integrated Assessment

The synthesis of multi-source evidence regarding the flood impact assessment in the Brazos River corridor following Hurricane Harvey reveals areas of strong agreement, particularly concerning the widespread nature of flooding and the high vulnerability of the affected population. Remote sensing analysis indicates a significant increase in flood severity from a pre-flood level of 1 to a post-flood level of 3, with a dominant fluvial overbank flooding pattern. This finding aligns with the socioeconomic vulnerability analysis, which identifies high vulnerability levels due to factors such as elevated road exposure, a substantial elderly and child population, and limited vehicle access for evacuation. Both analyses suggest that the flood has had a pronounced impact on the community's ability to respond effectively.

However, key uncertainties and conflicts arise in the details of exposure and vulnerability assessments. The street-level exposure analysis reports high road exposure but lacks specific data on building elevations and ground-floor exposure, which are critical for understanding the full extent of risk to individual structures. Additionally, the socioeconomic vulnerability analysis notes that aggregated indicators may not accurately reflect household-level vulnerabilities, suggesting a need for localized verification. The machine-learning flood risk prediction model indicates a high-risk score, but it is essential to recognize that this score is based on historical data and may not fully account for the unique circumstances of extreme events like Hurricane Harvey.

These uncertainties imply that while there is a consensus on the overall high risk and vulnerability, the specific impacts on buildings and households require further investigation. Human analysts should prioritize reviewing building elevations and ground-floor features to enhance the accuracy of exposure assessments. Additionally, local knowledge is crucial for contextualizing socioeconomic indicators and understanding community resources for flood response. Verification of the remote sensing data, particularly concerning water coverage and potential obstructions from cloud cover, is also necessary to ensure the reliability of the observed evidence.

Figure 6. Workflow of multi-source flood impact assessment as realized through the agentic framework (Input Images-Agent Prompts-Overall Assessment).

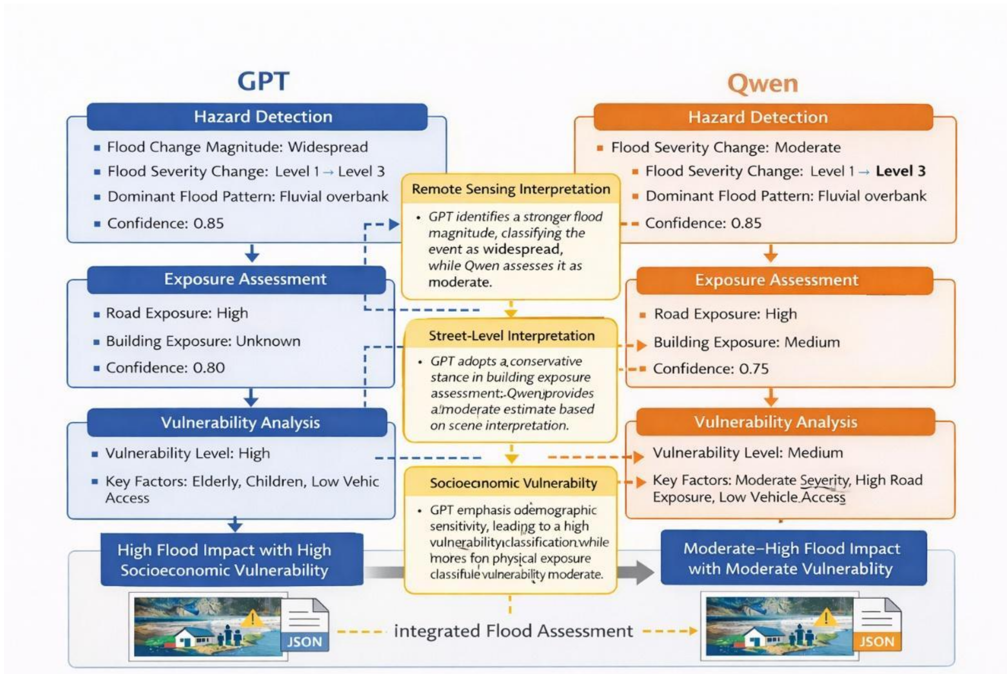


Figure 7. GPT vs Qwen Multimodal Flood Reasoning Comparison

Discussion and Conclusion

This study advances flood assessment by shifting from isolated hazard mapping toward a reliability-aware, human–AI collaborative spatial decision-support framework. By combining structured reasoning constraints, cross-agent verification, and multimodal evidence integration, the proposed architecture improves interpretive coherence across hazard, exposure, and vulnerability assessments.

The current study is limited by qualitative evaluation and a single Hurricane Harvey case study and should therefore be interpreted as an exploratory demonstration rather than systematic validation. Future work will incorporate ground-truth-based quantitative evaluation, additional flood events, and structured human review. Cross-source disagreement will also be examined to determine how evidence reliability varies by data source, spatial scale, and assessment task.

Overall, the proposed framework demonstrates how multimodal generative AI can move beyond descriptive capabilities toward structured spatial reasoning for hazard assessment. By integrating satellite imagery, street-level evidence, and socioeconomic context within an agentic orchestration architecture, the system provides a foundation for reliable and interpretable flood decision-support workflows.

Reference

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Zhu, T. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Clement, M. A., Kilsby, C. G., & Moore, P. (2018). Multi-temporal synthetic aperture radar flood mapping using change detection. *Journal of Flood Risk Management*, *11*(2), 152-168.

Hu, Y., Yuan, J., Wen, C., Lu, X., Liu, Y., & Li, X. (2025). Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, *224*, 272-286.

Huang, W., Wang, J., & Cong, G. (2024). Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science*, *38*(7), 1414-1442.

Kim, J., Jang, L., Chiang, Y. Y., Wang, G., & Pasco, M. C. (2025, November). StreetLens: Enabling Human-Centered AI Agents for Neighborhood Assessment from Street View Imagery. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Human-Centered Geospatial Computing* (pp. 15-19).

Kumbam, P. R., & Vejre, K. M. (2024). Floodlense: A framework for chatgpt-based real-time flood detection. *arXiv preprint arXiv:2401.15501*.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892-34916.

Liu, Z., Mohammadi, N., & Taylor, J. E. (2025). FloodVision: Urban Flood Depth Estimation Using Foundation Vision-Language Models and Domain Knowledge Graph. *arXiv preprint arXiv:2509.04772*.

Lu, K., Zhao, X., Li, M., Wang, Z., Zhou, Y., & Wang, J. (2025). StreetSenser: a novel approach to sensing street view via a fine-tuned multimodal large language model. *International Journal of Geographical Information Science*, 1-29.

McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International journal of remote sensing*, *17*(7), 1425-1432.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., & Murphy, R. R. (2021). Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, *9*, 89644-89654.

Schumann, G., & Di Baldassarre, G. (2010). The direct use of radar satellites for event-specific flood risk mapping. *Remote Sensing Letters*, *1*(2), 75-84.

Twele, A., Cao, W., Plank, S., & Martinis, S. (2016). Sentinel-1-based flood mapping: a fully automated processing chain. *International Journal of Remote Sensing*, *37*(13), 2990-3004.

Yang, Y., Xu, N., & Yang, J. J. (2025). Multi-agent visual-language reasoning for comprehensive highway scene understanding. *arXiv preprint arXiv:2508.17205*.

Zhan, Y., Xiong, Z., & Yuan, Y. (2025). Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, *221*, 64-77.

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, *5*(4), 8-36.

Zhu, J., Dang, P., Cao, Y., Lai, J., Guo, Y., Wang, P., & Li, W. (2024). A flood knowledge-constrained large language model interactable with GIS: enhancing public risk perception of floods. *International Journal of Geographical Information Science*, *38*(4), 603-625.