When Social Media Images Need Words: Measuring Context Gap and Fusion Tax in Crisis Image Captioning

Yuhao Bao
Brigham Young University
yyyyy@byu.edu

Amanda Lee Hughes
Brigham Young University
amanda_hughes@byu.edu

ABSTRACT

Crisis images on social media can be difficult to interpret at scale, and many carry meaning embedded in text or symbols (e.g., radar screenshots, evacuation notices). This limits vision-only captioning for situational awareness. We quantify a central trade-off in multimodal captioning: adding post text can reduce omission-driven ambiguity (the Context Gap), but it can also introduce text-driven errors (the Fusion Tax). Using 204 high-priority image–post pairs from CrisisFACTS, we compare Vision-only and Vision + Text captioning across Gemini 2.0 Flash, Qwen2.5-VL, and BLIP. We find that post text improves accuracy for Gemini and Qwen largely by reducing misidentification and scene-type errors, while sometimes amplifying hallucinated (unsupported) details. BLIP, however, does not reliably fuse modalities in our setup. When post text is provided, it often collapses into simple text echoing rather than producing image-grounded captions. We discuss implications for multimodal fusion in crisis informatics and outline next steps for image-type evaluation and routing.

Keywords

Crisis Informatics, Situational Awareness, Image Captioning, Multimodal Large Language Models, Social Media

INTRODUCTION

Social media platforms have become central channels through which affected individuals and emergency responders exchange information and support time-sensitive decisions during disaster events (Vieweg et al., 2010; Palen & Hughes, 2018). People caught in crises turn to platforms such as X (formerly Twitter) and Facebook to document conditions, warn others, and coordinate safety decisions (Hughes & Palen, 2009; Vieweg et al., 2010). In parallel, emergency managers attend to social media streams to support situational awareness and response decision-making (Reuter et al., 2018; Zade et al., 2018). Yet the scale and heterogeneity of social media data make it difficult to locate information that is both relevant and actionable under time pressure and surge volumes (Imran et al., 2015; Purohit et al., 2025).

This challenge is especially pronounced for visual content. Social media images can provide direct evidence of damage severity, flood boundaries, fire conditions, and evacuation activity (Nguyen et al., 2017; Alam et al., 2018; Mouzannar et al., 2018), but they are costly to triage and interpret at scale. Captioning offers a lightweight initial filter for rapid assessment and human handoff, without requiring event-specific training labels. Multimodal large language models (MLLMs) present a promising path forward because they can generate crisis-relevant image captions rapidly and at scale (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). However, crisis captioning is only operationally useful if it captures what happened, where it happened, and what action is needed to support decision-making, rather than generating generic scene descriptions. Our pilot analysis revealed two recurring failure modes. The *Context Gap* arises when vision-only models produce incomplete or ambiguous captions because the situational context needed to resolve meaning is absent. The *Fusion Tax* arises when post text is incorporated but overweighted relative to the image, introducing text-driven errors that are not visually grounded. The central question, then, is not whether context helps in general, but when closing the Context Gap is worth

paying the Fusion Tax.

This paper makes three contributions. First, we provide a baseline empirical comparison of Vision-only and Vision + Text captioning for high-priority crisis imagery, isolating the effect of textual context across three models (Gemini 2.0 Flash, Qwen2.5-VL, and BLIP). Second, we introduce an evaluation protocol and error taxonomy that distinguishes omission-driven Context Gap errors from text-induced Fusion Tax errors. Finally, we outline operational implications for crisis-response pipelines by identifying when post text improves caption quality, when it introduces grounding risks, and why image type and native text-in-image capabilities matter for deployment.

BACKGROUND

Crisis informatics research has often centered on text streams, in part because text is comparatively straightforward to collect and analyze at scale and it supports core response-oriented tasks such as event detection, situational summarization, and surfacing emergency information needs (Imran et al., 2015; Reuter et al., 2018). However, a primarily text-centered approach can overlook evidence that is expressed visually, or only weakly represented in accompanying text.

To address this gap, prior work has explored crisis image processing through labeling, classification, and damage assessment. These approaches can enable filtering pipelines and rapid categorization (e.g., infrastructure damage versus unaffected scenes), but they often depend on supervised learning and labeled datasets (Alam et al., 2018; Nguyen et al., 2017; Mouzannar et al., 2018). That reliance is challenging in unfolding events, where visual patterns shift, labels may not align with emerging realities, and key cues may be subtle or context dependent—creating familiar limits around generalization and domain shift (Li et al., 2019; Khattar & Quadri, 2022).

Recent advances in large language models (LLMs) and multimodal large language models (MLLMs) introduce new options for interpreting crisis-related content (Clark & Hughes, 2025). LLM-based systems can summarize large volumes of noisy information, but they also raise reliability concerns, including susceptibility to misinformation, prompt sensitivity, and fluent outputs that are not well supported by evidence (Ji et al., 2023; Huang et al., 2025). For visual content, MLLMs can process images jointly with associated text, which enables captioning and interpretation that may be more flexible than task-specific classifiers or conventional captioning pipelines (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Clark & Hughes, 2025). These capabilities motivate careful evaluation in crisis settings. If a model makes an error, that mistake can be carried into downstream filtering and decision-support tools and influence response decisions.

A related challenge is the prevalence of what we term information-dense crisis imagery. We define this category as visual artifacts in which operationally relevant semantics are conveyed primarily through embedded text, legends, symbols, or graphical encodings rather than through naturalistic scene content. We operationalize this classification by asking whether removing on-image textual or symbolic overlays would eliminate the primary situational meaning of the image. Images in which key operational signals are encoded in labels, legends, icons, geographic markers, or domain-specific symbology are classified as information-dense.

Examples include Doppler radar warning maps, evacuation notices, official alert flyers, and crisis-related screenshots. In such images, the critical signal for situational awareness is carried by structured visual encodings—such as labels, color-coded regions, and annotated overlays—rather than by recognizable real-world objects (see Figure 1). In the high-priority subset analyzed in this study (N = 204), 173 images (84.8%) meet this operational criterion. Accordingly, caption quality may depend on both multimodal fusion and native visual encoding capabilities, including OCR and fine-grained recognition of labels, legends, and map symbology. Crisis social media is heavily populated by such information-dense visuals (Prestley & Morss, 2023), which makes robust text-in-image understanding a critical capability for operational captioning (Singh et al., 2019; Mishra et al., 2019; Mathew et al., 2021).

These dynamics suggest that caption performance in crisis settings depends jointly on (1) visual encoding of information-dense artifacts and (2) the reliability of multimodal fusion. This motivates the structured comparison of Vision-only and Vision + Text captioning that we take up in the sections that follow.

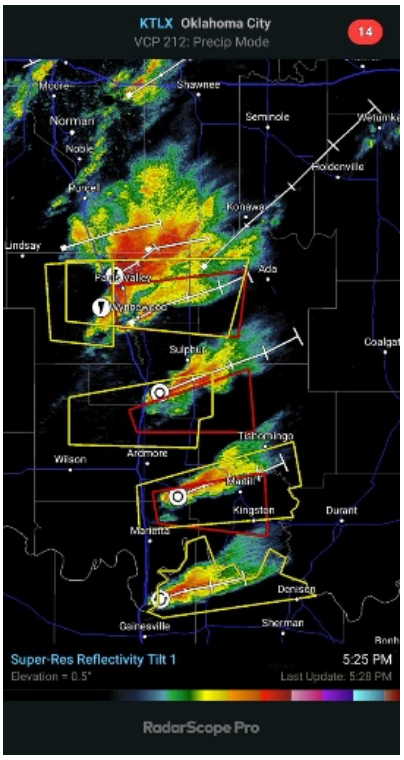

Figure 1. Example of information-dense crisis imagery from the CrisisFACTS dataset

METHODOLOGY

This study evaluates how adding accompanying post text affects the operational accuracy of crisis image captions produced by current vision–language captioning systems. We compare two input conditions—Vision-only (image only) and Vision + Text (image plus its associated post text)—and manually score caption accuracy and failure modes using a structured rubric and error taxonomy (Tables 2–3). The design is intended to isolate the effect of textual context on caption quality and characterize the trade-off between omission-driven failures (Context Gap) and text-driven errors introduced or amplified through multimodal fusion (Fusion Tax).

Data Collection

We use a subset of the CrisisFACTS benchmark (McCreadie & Buntain, 2023; Buntain et al., 2023), developed within the TREC-IS initiative to support evaluation of situational-awareness extraction from crisis-related social media. We focus on image–post pairs from Twitter (now X), where each image is paired with its associated post text. Our sample consists of N = 204 image-post pairs drawn from the High-Priority Images subset. These pairs were labeled by human coders as high or critical priority for responder situational awareness (i.e., judged to contain actionable crisis-relevant information such as visible damage, hazards, or response guidance). The sample spans multiple disaster events (e.g., the 2020 Beirut Explosion, tornado events in Oklahoma, and hurricane events such as Hurricane Florence and Laura) and includes both naturalistic photographs and information-dense artifacts (e.g., radar maps and evacuation notices).

Model Selection

To assess when the benefits of post text outweigh the risks of fusion-driven errors, we compare the two captioning conditions across three model architectures. We evaluate three captioning approaches that reflect practical options for crisis-response pipelines: BLIP (Li et al., 2022) as a non-LLM image-captioning baseline, Qwen2.5-VL (Bai et al., 2025) as an open-source MLLM, and Gemini 2.0 Flash (Google DeepMind, 2025) as a commercially available MLLM. This mix supports comparisons across (i) conventional captioning versus MLLMs and (ii) commercial versus open-source deployment pathways relevant to tool-building and reproducible workflows.

The BLIP image-captioning-large model provides a cost-effective baseline. Gemini 2.0 Flash serves as the primary reference model given its strong multimodal performance and practical deployment characteristics (e.g., speed and cost profile for large-scale captioning). Qwen2.5-VL is included to represent an open-source MLLM with enhanced native visual comprehension (including text-in-image understanding), enabling comparison not only between Vision-only and Vision + Text, but also across architectures that may reduce dependence on external post text for artifact-heavy image classes.

Prompting and Experimental Conditions

To isolate the impact of post text, we generate captions under two conditions:

1. **Vision-only:** image input only
2. **Vision + Text**: image plus its associated post text

Prompts were adapted to each model's interface while preserving a shared goal: produce a concise, crisis-relevant caption grounded in visible evidence (Table 1). For Gemini and Qwen, we use a consistent zero-shot instruction specifying a single-sentence caption with no introductory phrases and no speculative claims. For BLIP, we use a standard captioning prompt for Vision-only, and a more structured format in the Vision + Text condition to reduce instruction/text echoing when post text is supplied.

Across all models and conditions, captions are constrained to one sentence to support comparability and reflect an operational requirement for scannable outputs. Table 1 summarizes the prompt structure by model and condition.

Table 1. Prompt Specifications by Model and Input Condition

Model	Input	Prompt Structure
Gemini 2.0 Flash	Vision-only	One-sentence caption; no introductory phrases.
	Vision + Text	One-sentence caption integrating post text and image; include actor/object, action/state, setting, and visible signals; no speculation.
Qwen2.5-VL	Vision-only	Same as Gemini (Vision-only).
	Vision + Text	Same as Gemini (Vision + Text).
BLIP	Vision-only	Standard zero-shot: "An image of..."
	Vision + Text	Image + structured prompt: "Context: [Tweet_Text] Question: What is happening in this image? Answer:"

Evaluation

We developed the accuracy rubric and error taxonomy through iterative coding. We began with a pilot review of 30 randomly selected images to identify recurring failure modes, which informed our definitions of Context Gap and Fusion Tax. Generated captions were then manually evaluated using a three-level accuracy rubric (Table 2) and an error taxonomy (Table 3). The lead author performed the primary coding, with periodic consultation with the second author to refine the codebook and resolve ambiguous cases. Our primary metric is Accuracy Rate, defined as the proportion of captions labeled Accurate.

To assess coding reliability, the second author independently coded a random subset of n = 100 captions using the same three-level accuracy rubric. Interrater reliability was Krippendorff's α (nominal) = 0.69 (percent agreement = 93.0%). Remaining disagreements were limited to Accurate vs. Partially Accurate and were resolved by discussion; adjudicated labels were used in analysis.

For captions labeled Partially Accurate or Inaccurate, we assign a primary failure mode within one of two error families. Context Gap errors capture omission- or misinterpretation-driven failures attributable to insufficient visual grounding (e.g., missing key situational cues, misidentifying the scene type). Fusion Tax errors capture failures introduced or amplified when post text is incorporated, including image-incongruent hallucinations and prompt/text echoing.

Table 2. Caption Accuracy Rubric

Level	Definition
Accurate	Caption accurately captures the key hazard, setting, and severity without errors or critical omissions.
Partially Accurate	Caption captures the gist but omits crisis-relevant specifics and/or contains minor inaccuracies that do not materially change interpretation.
Inaccurate	Caption is unrelated to the image, includes major unsupported details, or misidentifies crisis type/setting in a way that materially changes interpretation.

Table 3. Error Taxonomy

Error Family	Subcategory	Definition
Context Gap	*Key Element Omission*	Missing primary hazard or operationally critical element.
	Scene Type Error	Misinterpreting the broader event context.
	Object Misidentification	Incorrectly identifying key objects.
Fusion Tax	*Hallucination*	Introducing details not supported by the image (often amplified when post text is incorporated).
	Prompt/Text Echoing	Verbatim or near-verbatim repetition of the prompt and/or input post text, producing a non-informative caption.
	Event Causality Error	Inferring causal explanations not supported by the image (often text-driven)

RESULTS

We evaluated six model–input conditions on N = 204 crisis-relevant image-post pairs: BLIP, Gemini 2.0 Flash, and Qwen2.5-VL, each run in Vision-only and Vision + Text modes. Five conditions produced scorable, image-grounded captions and were rated using the rubric in Table 2. BLIP (Vision + Text) is reported only as a failure case: all 204 outputs echoed the prompt or post text and therefore did not satisfy the caption criterion.

In Vision-only mode, BLIP achieved an Accurate rate of 49.5% (101/204). For the two MLLMs, adding post text improved accuracy: Gemini 2.0 Flash increased from 75.5% (154/204) to 91.7% (187/204) (+16.2 points), and Qwen2.5-VL increased from 77.0% (157/204) to 86.8% (177/204), a 9.8-point improvement. Overall, both models benefit from textual context, with a larger gain for Gemini.

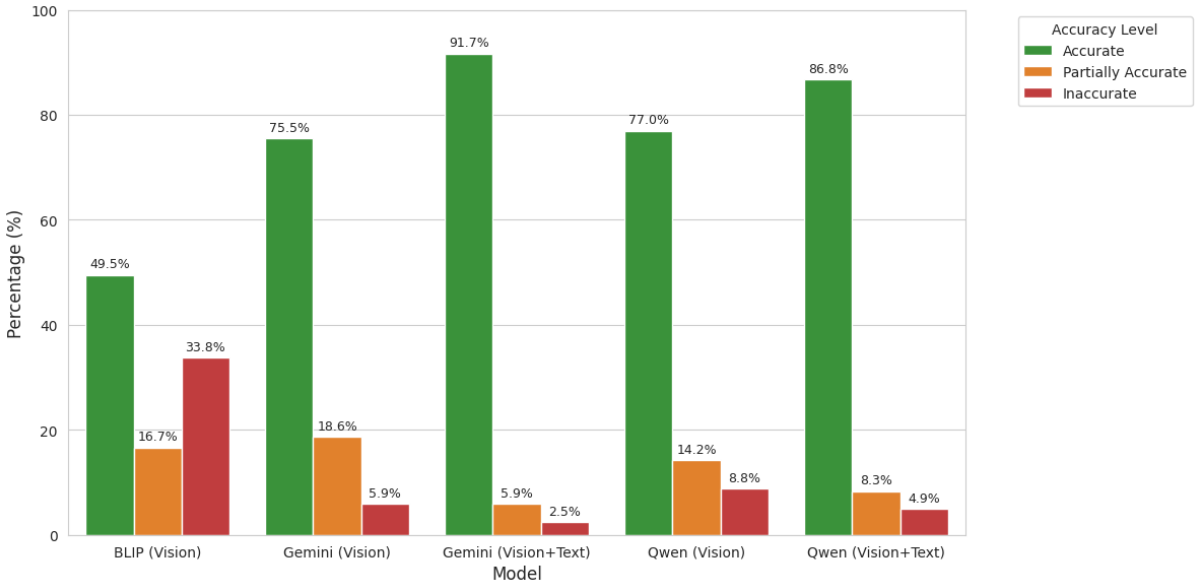

Figure 2: Caption Accuracy Across Models and Modalities (Vision vs. Vision + Text)

Error Analysis: Context Gap vs. Fusion Tax

To understand why captions fail under each condition, we coded non-Accurate outputs using the Context Gap vs. Fusion Tax taxonomy (Table 3). Table 4 reports the distribution of primary failure modes by model and modality. Failure modes reflect the primary error code assigned to captions rated Partially Accurate or Inaccurate; captions rated Accurate receive no error code. For completeness, BLIP (Vision + Text) appears in Table 4 as an unscored failure case, but it is excluded from accuracy comparisons and Figure 2.

In the Vision-only setting, most errors reflected the Context Gap, as models often lacked the contextual grounding needed to interpret ambiguous scenes and information-dense visuals. Gemini most often failed via Object Misidentification (32/204; 15.7%), followed by Scene Type Error (13/204; 6.4%) and Key Element Omission (5/204; 2.5%). Qwen2.5-VL showed a similar pattern, with its non-Accurate captions also dominated by Context Gap codes in Table 4, including object- and scene-level misinterpretations in cases where the visual evidence is hard to decode without context. For example, Doppler radar images showing severe weather systems were sometimes described as generic "precipitation maps," completely missing the critical Tornado Warning context. Similarly, images from the Beirut explosion were occasionally described as "clouds" or "industrial smoke," failing to classify the event as an explosion. Furthermore, models struggled with Object Misidentification, such as misclassifying specialized emergency response vehicles as generic commercial trucks due to a lack of situational clues.

In the Vision + Text condition, adding post text reduced Context Gap errors for both models but introduced Fusion Tax failures. For instance, Gemini's object misidentification dropped from 32/204 (15.7%) to 6/204 (2.9%), and Qwen shows the same directional shift in Table 4, with fewer context-driven mistakes once post text supplies event cues and labels. Vision + Text also introduced Fusion Tax failures. In particular, text-driven hallucinations emerged when the model incorporated details suggested by the post but not supported by the image (e.g., inferring a "stage setting" from mention of a press conference). In other cases, disaster keywords in the post led models to miss satirical or non-crisis framing and generate captions that were too serious relative to the visual content. Hallucination rates were low for Gemini (3/204; 1.5%) but higher for Qwen2.5-VL (8/204; 3.9%), suggesting that the benefits of post text can come with greater susceptibility to linguistically plausible—

but visually ungrounded—overreach, depending on model fusion behavior.

Table 4. Distribution of Error Types by Model and Input Condition. Percentages are computed over the full sample (N = 204). Failure modes reflect the primary error code for non-Accurate captions.

Model	Input	Error Family	Failure Mode	Count	%
BLIP	Vision	Context Gap	Scene Type Error	68	33.3
			Key Element Omission	27	13.2
			Object Misidentification	8	3.9
	Vision + Text	Fusion Tax	Prompt/Text Echoing	204	100.0
Gemini 2.0 Flash	Vision	Context Gap	Object Misidentification	32	15.7
			Scene Type Error	13	6.4
			Key Element Omission	5	2.5
	Vision + Text	Context Gap	Object Misidentification	6	2.9
			Scene Type Error	6	2.9
		Fusion Tax	Hallucination	3	1.5
			Event Causality Error	2	1.0
Qwen2.5-VL	Vision	Context Gap	Object Misidentification	21	10.3
			Scene Type Error	15	7.4
			Key Element Omission	7	3.4
		Fusion Tax	Hallucination	4	2.0
	Vision + Text	Context Gap	Scene Type Error	7	3.4
			Object Misidentification	5	2.5
			Key Element Omission	2	1.0
		Fusion Tax	Hallucination	8	3.9
			Event Causality Error	3	1.5
			Prompt/Text Echoing	2	1.0

Exploring Information-Dense Images and Native OCR

Of the 204 image-post pairs in the high-priority subset, 173 (84.8%) were classified as information-dense according to the operational definition described above. This indicates that responder-relevant content in this subset is heavily artifact-oriented, with a large proportion of maps, warning graphics, and official screenshots. As a result, caption performance here is less a test of natural image description and more a test of text-in-image understanding. We also examined whether models with stronger native visual processing—particularly OCR—can extract operational detail directly from information-dense visuals such as maps and warning products. These examples are illustrative; we do not separately score OCR extraction quality in this study.

Qualitative analysis revealed systematic differences across models. For instance, in a map-based visualization associated with the Beirut explosion (see Figure 3), models captioned a crisis-related map with embedded textual indicators.

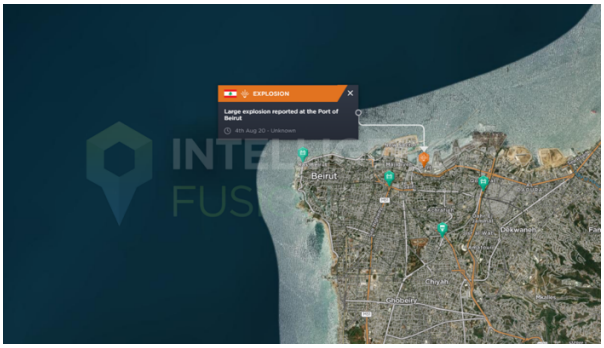

Figure 3. Map-based visualization associated with the Beirut port explosion.

BLIP produced a generic description of a city map without identifying the crisis context. Gemini described the map structure but failed to recognize key textual indicators referencing the explosion location. In contrast, Qwen2.5-VL incorporated text embedded in the visualization and generated a caption that correctly referenced the Beirut port explosion, producing a more situationally informative description.

A similar pattern appeared in severe weather warning maps that combine geographic regions with operational text and hazard indicators. When captioning the image found in Figure 4, BLIP generated high-level descriptions of a weather map without operational details. Gemini identified the presence of severe weather conditions but did not consistently reference labeled regions. Qwen2.5-VL successfully extracted warning information and geographic references visible within the image, indicating stronger text-in-image understanding.

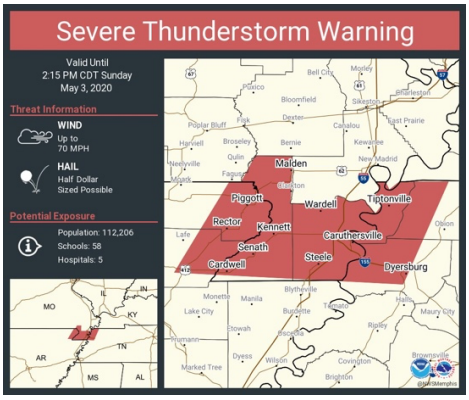

Figure 4. Severe thunderstorm warning graphic with embedded operational details.

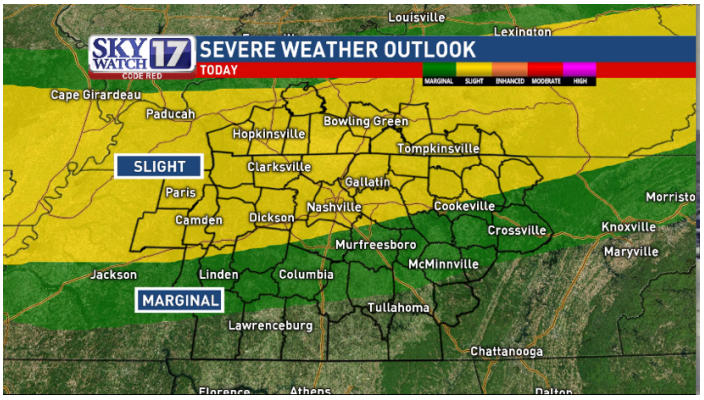

Figure 5. Severe weather outlook risk map illustrating labeled threat categories.

In a severe weather outlook visualization containing labeled risk categories and color-coded threat levels, key semantics are embedded directly within the image. When captioning the image in Figure 5, Qwen2.5-VL again incorporated warning status information directly from image text, whereas BLIP and Gemini primarily relied on generalized visual interpretation. These examples suggest that stronger native visual encoding—particularly text-in-image (OCR) capability—can reduce reliance on external post text when interpreting information-dense artifacts. This finding motivates future work on modality-aware routing strategies, such as directing map- or flyer-like images to OCR-capable captioning pipelines.

DISCUSSION

These findings show that the Context Gap–Fusion Tax trade-off is not uniform: it shifts with model architecture and with what the model can reliably extract from the image. In our sample, adding post text improved accuracy for both MLLMs largely by reducing Context Gap errors, but it also introduced a smaller set of Fusion Tax failures when linguistic cues were overweighted relative to visual evidence. While we do not propose or evaluate routing policies here, the patterns we observe point toward modality-aware pipeline strategies as a concrete next design direction.

Closing the Context Gap Comes with a Fusion Tax

The clearest operational implication is that post text helps most when the image alone does not carry enough situational grounding to support crisis-relevant interpretation. In the Vision-only condition, errors were dominated by Context Gap failure modes: object misidentification, scene-type errors, and omission of key hazard details. This pattern aligns with a basic feature of crisis imagery in the wild. Operationally, the key issue is often the meaning of an image in context, not just its visible contents (e.g., whether smoke indicates routine industrial activity versus an explosion aftermath).

In some cases, text fusion can introduce new failure modes when linguistic cues pull the caption away from what is visually supported, a failure mode documented in recent work on MLLM reliability (Ji et al., 2023; Chen et al., 2024). In our data, this showed up as hallucinated details: attributes implied by the post but not present in the image. We also observed forced crisis framings when disaster keywords appeared alongside satire or non-crisis imagery. This is the Fusion Tax: gains in completeness and specificity can be purchased with new risks to grounding.

For crisis workflows, this means that captioning should not be treated as uniformly reliable across posts. A more

operational stance is to design systems that surface how a caption was formed—whether it is primarily image-grounded versus text-driven—and that support triage under uncertainty (e.g., flagging captions that depend heavily on post text, or highlighting which caption elements are directly evidenced in the image).

Information-Dense Images and the Role of Native OCR

A second theme is the prevalence of information-dense crisis imagery—radar maps, evacuation notices, warning flyers, and screenshots—where key semantics are carried through embedded text, symbols, legends, and graphical encodings. For these images, generic scene description often fails to recover actionable content because the "signal" is not the scene; it is the labeled overlay. Our qualitative examples suggest that models with stronger text-in-image understanding can sometimes extract operational detail directly from these artifacts, potentially reducing reliance on external post text for certain image classes. Because we did not score OCR extraction quality separately, these observations should be read as qualitative indicators rather than an OCR benchmark. More broadly, the prevalence of information-dense imagery reinforces that a single captioning strategy is unlikely to be dependable across the full range of crisis visuals. This observation motivates future work on modality-aware routing (e.g., directing map/flyer/screenshot-like images to OCR-capable pipelines, and using Vision + Text for ambiguous naturalistic images with safeguards against text-driven overreach).

Model Choice as a Reliability–Cost Trade-Off

Model choice shaped both baseline performance and susceptibility to Fusion Tax errors. Gemini achieved the highest accuracy when paired with post text, while Qwen2.5-VL showed higher hallucination rates under text fusion, though the absolute counts were small in this sample. This suggests different sensitivities to linguistic noise and different calibration needs.

Deploying these models in crisis environments means balancing accuracy against cost, latency, and privacy constraints (such as the risks of third-party APIs). The operational goal is not a perfect description. Models instead need to provide reliable baselines for rapid triage and sense-making under time pressure (Hiltz & Plotnick, 2013; Reuter et al., 2018). When automated filtering successfully reduces visual noise, response teams can assess conditions faster and prioritize verified observations over unverified reports. From a deployment perspective, our Gemini 2.0 Flash runs for this study were low-cost[1] under the pricing available at the time of data collection.

Limitation and Future Work

This study has several limitations. The 204-pair sample limits generalizability across diverse crisis events, languages, and platforms. Furthermore, because we evaluate the baseline utility of off-the-shelf models rather than proposing new architectures, we do not include formal ablation studies or statistical significance testing. Assessing "partial accuracy" also remains difficult; correct interpretation often depends on what an emergency responder would reasonably infer in the moment. Finally, our text-fusion condition assumes the accompanying social media post is factual. In practice, online text is frequently noisy, sarcastic, or deceptive, which likely worsens fusion errors.

Future work will address these gaps. Expanding the dataset across more events will enable proper statistical validation. We also need to test model robustness against misleading text and analyze performance variations across specific image classes. Before developing custom models, an important next step is to evaluate prompting strategies and fusion mechanisms that keep outputs strictly grounded in visual evidence. Testing safeguards like uncertainty flags—particularly across different types of visual artifacts—will help clarify the requirements for modality-aware routing in operational pipelines.

CONCLUSION

This paper evaluates multimodal captioning for crisis-related social media imagery using a high-priority subset of CrisisFACTS and a manual rubric that attends to both correctness and how captions go wrong. Across the two contemporary vision–language models evaluated (Gemini & Qwen2.5-VL), incorporating post text generally improves caption accuracy by supplying missing situational grounding. At the same time, it introduces a smaller—but meaningful—class of fusion-driven errors when linguistic cues are over-weighted or misleading. Meanwhile, the BLIP baseline under Vision + Text collapses into prompt/text echoing, underscoring that fusion must be handled carefully. Our taxonomy separates Context Gap and Fusion Tax failure modes and points to several design implications: treating image types differently, leveraging strong OCR for information-dense artifacts, and surfacing uncertainty and evidentiary grounding in caption outputs. As MLLMs become more integrated into emergency management workflows, understanding when and how to trust their outputs becomes as important as improving their accuracy.

[1] Approximately $0.10 USD.

REFERENCES

Alam, F., Ofli, F., & Imran, M. (2018). CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.14983

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, *35*, 23716-23736.

Bai, S., Chen, K., Liu, X., Wang, J., et al. (2025). Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*. https://doi.org/10.48550/arXiv.2502.13923

Buntain, C., Hughes, A. L., McCreadie, R., Horne, B. D., Imran, M., & Purohit, H. (2023). CrisisFACTS 2023-Overview Paper. In Proceedings of the TREC 2023 Conference. https://trec.nist.gov/pubs/trec32/papers/Overview_crisis.pdf

Clark, H., & Hughes, A. L. (2025). Seeing the Storm: Leveraging Multimodal LLMs for Disaster Social Media Video Filtering. Proceedings of the International ISCRAM Conference. https://doi.org/10.59297/f9bnkx60

Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., ... & Chen, H. (2024). Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3235–3252). https://aclanthology.org/2024.acl-long.178/

Google DeepMind. (2025). *Gemini 2.0 Flash* [Model description]. Retrieved from https://deepmind.google/models/gemini/flash/

Starr Roxanne Hiltz, & Linda Plotnick. (2013). Dealing with information overload when using social media for emergency management: Emerging solutions. In *ISCRAM 2013 Conference Proceedings – 10th International Conference on Information Systems for Crisis Response and Management* (pp. 823–827). KIT; Baden-Baden: Karlsruher Institut fur Technologie. https://idl.iscram.org/files/hiltz/2013/583_Hiltz+Plotnick2013.pdf

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Article 42. https://doi.org/10.1145/3703155

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3/4), 248–260. https://doi.org/10.1504/IJEM.2009.031564

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, 47(4), Article 67. https://doi.org/10.1145/2771588

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. https://doi.org/10.1145/3571730

Khattar, A., & Quadri, S. M. K. (2022). Generalization of convolutional network to domain adaptation network for classification of disaster images on Twitter. *Multimedia Tools and Applications*, 81, 30437–30464. https://doi.org/10.1007/s11042-022-12869-1

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888-12900). PMLR.

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.

Li, X., Caragea, D., Caragea, C., Imran, M., & Ofli, F. (2019). Identifying Disaster Damage Images Using a Domain Adaptation Approach. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management*. https://idl.iscram.org/files/xukunli/2019/1853_XukunLi_etal2019.pdf

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, *36*, 34892-34916.

Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). Docvqa: A dataset for vqa on document images.

In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2200-2209).

McCreadie, R., & Buntain, C. (2023). CrisisFACTS: Building and Evaluating Crisis Timelines. In *Proceedings of the 20th International ISCRAM Conference* (pp. 320–339). Omaha, USA: University of Nebraska at Omaha. http://dx.doi.org/10.59297/JVQZ9405

Mishra, A., Shekhar, S., Singh, A. K., & Chakraborty, A. (2019, September). Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)* (pp. 947-952). IEEE.

Mouzannar, H., Rizk, Y. & Awad, M (2018). Damage Identification in Social Media Posts using Multimodal Deep Learning. In *ISCRAM 2018 Conference Proceedings – 15th International Conference on Information Systems for Crisis Response and Management* (pp. 529–543). Rochester, NY (USA): Rochester Institute of Technology. http://idl.iscram.org/files/husseinmouzannar/2018/2129_HusseinMouzannar_etal2018.pdf

Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 569–576. https://doi.org/10.1145/3110025.3110109

Palen, L., & Hughes, A. L. (2018). Social Media in Disaster Communication. In H. Rodriguez, W. Donner, & J. E. Trainor (Eds.), *Handbook of Disaster Research* (pp. 497–518). Springer International Publishing. https://doi.org/10.1007/978-3-319-63254-4_24

Prestley, R., & Morss, R. E. (2023). Contextualizing disaster phases using social media data: Hurricane risk visualizations during the forecast and warning phase of Hurricane Irma. *Weather, climate, and society*, *15*(4), 1049-1067.

Purohit, H., Buntain, C., Hughes, A. L., Peterson, S., Lorini, V., & Castillo, C. (2025). Engage and Mobilize! Understanding Evolving Patterns of Social Media Usage in Emergency Management. *Proceedings of the ACM on Human-Computer Interaction, (CSCW)*. https://doi.org/10.1145/3710965

Reuter, C., Hughes, A. L., & Kaufhold, M.-A. (2018). Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of Human–Computer Interaction*, 34(4), 280–294. https://doi.org/10.1080/10447318.2018.1427832

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., ... & Rohrbach, M. (2019). Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8317-8326).

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, 1079–1088. https://doi.org/10.1145/1753326.1753486

Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., & Starbird, K. (2018). From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response. Proceedings of the ACM Human-Computer Interaction, 2(CSCW), 195:1-195:18. https://doi.org/10.1145/3274464