VLCE: An External Knowledge Framework for Contextual Image Captioning in Disaster Assessment

Md. Mahfuzur Rahman[1], Marufa Kamal[2], Fahad Rahman[3], Sunzida Siddique[4], Ahmed Rafi Hasan[3], Mohd Ariful Haque[1], Kishor Datta Gupta[1], Roy George[1]

[1]Clark Atlanta University, Atlanta, GA 30314, USA,
[2]BRAC University, Dhaka 1212, Bangladesh,
[3]United International University, Dhaka 1212, Bangladesh,
[4]Daffodil International University, Dhaka, Bangladesh

Correspondence: mdmahfuzur.rahman@students.cau.edu

Abstract

General-purpose vision-language models (VLMs) such as LLaVA and QwenVL produce descriptions of disaster imagery that lack domain-specific vocabulary and actionable detail. We propose the Vision-Language Caption Enhancer (VLCE), a framework that integrates external semantic knowledge from ConceptNet and WordNet into the caption generation process for post-disaster satellite and UAV imagery. VLCE operates in two stages: first, a baseline VLM generates an initial caption conditioned on YOLOv8 object detections; second, a knowledge-enriched sequential model, a CNN-LSTM or a hierarchical cross-modal Transformer, refines the caption using a vocabulary augmented with 1,566 domain-relevant terms extracted from knowledge graphs. We evaluate VLCE on two disaster benchmarks: xBD (satellite, 6,369 images, 3 damage classes) and RescueNet (UAV, 4,494 images, 12 damage classes), using CLIPScore for semantic alignment and InfoMetIC for informativeness. On RescueNet with the Transformer decoder, VLCE with knowledge graph enrichment produces captions preferred over QwenVL baselines in 95.33% of image pairs on InfoMetIC and 73.64% on CLIPScore. Qualitative analysis shows that without knowledge graph integration, generated captions exhibit hallucinations, word repetition, and semantic incoherence, whereas knowledge-enriched captions maintain factual consistency and domain-appropriate vocabulary. intended as a continuous, extensible monitor of differential framing under changing real-world inputs.

1 Introduction

Natural disasters cause widespread destruction to infrastructure, ecosystems, and communities. Rapid and accurate assessment of post-disaster damage is critical for coordinating emergency response, allocating resources, and planning recovery efforts. Satellite and unmanned aerial vehicle (UAV) images are now the main sources of data for disaster assessment. They give clear pictures of large areas of land.

Figure 1: Post-hurricane disaster imagery from the RescueNet dataset.

Recent advances in vision-language models (VLMs) have enabled the automatic generation of textual descriptions from images. Models such as LLaVA (Liu et al., 2023) and QwenVL (Bai et al., 2023) achieve strong performance on general image captioning benchmarks. However, when applied to post-disaster imagery, these models produce generic descriptions, lack domain-specific terminology, and fail to capture the nuanced damage indicators that responders require. A caption stating "buildings in a residential area" provides little actionable information compared to one identifying "severely damaged roof structures with debris fields and evidence of flooding in surrounding roadways."

This gap arises because general-purpose VLMs are trained mainly on everyday imagery and lack exposure to the specialized vocabulary and visual

patterns characteristic of disaster scenarios (Mason and Charniak, 2014). The value of contextual captioning is illustrated in Figure 1. A model without domain knowledge produces: *"A satellite image depicts a community from an aerial perspective, revealing multiple dwellings and streets"* In contrast, a knowledge enriched model generates: *"The image shows the aftermath of Hurricane Michael, which inflicted major damage to infrastructure and the environment. Roads and streets are cluttered with debris from destroyed structures, such as fallen trees and scattered items. Trees in the region have suffered varied degrees of damage, with some seeming damaged or totally fallen. Buildings show different degrees of destruction, ranging from comparatively intact ones to those with obvious traces of catastrophic damage. The entire scene depicts the ongoing recovery and rebuilding activities in the aftermath of the catastrophe"*, defining the type of disaster, outlining the effects on infrastructure, and communicating the state of recovery. Adding domain-relevant outside knowledge to the caption generation pipeline is necessary to overcome this constraint.

1.1 Our Contributions

We propose VLCE (Vision-Language Caption Enhancer), a framework that enriches disaster image captioning through external knowledge graph integration. Our contributions are as follows:

1. **Knowledge-enriched captioning framework.** We introduce a two-stage pipeline that combines baseline VLM captions with knowledge graph-augmented sequential models. The system extracts keywords from training captions using RAKE, looks up related terms in ConceptNet and WordNet, and builds a vocabulary of 3,195 words (1,566 from knowledge graphs) that helps create more accurate descriptions of disasters.

2. **Dual architecture for satellite and UAV imagery.** We create two different decoder designs: one is a CNN-LSTM model that combines visual and text information, and the other is a hierarchical cross-modal Transformer that processes visual information. Each architecture is paired with a domain-specific image encoder: ResNet50-EuroSAT for satellite imagery and a ViT-based UAV classifier for drone imagery.

3. **Comprehensive evaluation on disaster benchmarks.** We evaluate xBD and RescueNet using CLIPScore and InfoMetIC, which are two complementary metrics that together capture both semantic alignment and informational content. We show that knowledge graph integration is essential, and without it, the quality of captions for UAV images drops significantly (less than 2% preference on both metrics for RescueNet), but with it, the Transformer model gets up to 95.33% preference compared to standard VLMs on InfoMetIC.

2 Related Work

Satellite-based damage assessment has progressed from manual interpretation to automated deep learning pipelines. The xBD dataset (Gupta et al., 2019) established a large-scale benchmark for building damage classification from satellite imagery, while RescueNet (Rahnemoonfar et al., 2023) extended coverage to UAV-captured scenes with fine-grained damage categories. Alisjahbana et al. (Alisjahbana et al., 2024) proposed a two-step CNN for building damage segmentation (0.66 F1). Abbas and Dang (Abbas and Dang, 2023) investigated CNN-based caption generation for disaster areas (BLEU-1: 0.8731, CIDEr: 5.0908), and Chun et al. (Chun et al., 2022) applied deep learning for bridge damage descriptions (92.9% accuracy). Multimodal approaches combining image and text data have shown promise for disaster classification (Zou et al., 2021; Kota et al., 2022). Most existing methods focus on damage *classification* rather than free-form description generation.

LLaVA (Liu et al., 2023) pairs a CLIP ViT-L/14 encoder with a LLaMA decoder through visual instruction tuning. QwenVL (Bai et al., 2023) extends the Qwen language model with a vision encoder for multi-image understanding. Cornia et al. (Cornia et al., 2020) explored meshed-memory Transformers for captioning (CIDEr 132.7 on MS COCO). Chen and Li (Chen and Li, 2025) introduced multi-modal graph aggregation Transformers. While these models excel on general benchmarks, their application to specialized domains such as disaster assessment remains limited by training data distribution (Mason and Charniak, 2014).

External knowledge bases have been used to enhance language understanding across a range of tasks.

ConceptNet (Schön et al., 2019) provides commonsense relational knowledge, while WordNet offers lexical relationships including synonymy and hypernymy. ConceptNet Numberbatch embeddings combine distributional semantics with structured knowledge, producing word vectors that encode both corpus statistics and relational information. Zhou et al. (Zhou et al., 2019) applied knowledge graphs to captioning on MS COCO and Visual Genome. Zhao and Wu (Zhao and Wu, 2024) proposed multi-modal KG methods for graph attention captioning. Tang et al. (Tang et al., 2022) introduced KG-guided attention for CNN-LSTM models. Wang et al. (Wang et al., 2024) explored KG strategies for remote sensing captioning. Despite these advances, KG application to disaster-specific imagery has not been explored.

Pre-trained word embeddings serve as the bridge between discrete vocabulary tokens and continuous neural network inputs. Comparative analyses (Toshevska et al., 2020; Elbedwehy et al., 2023) have evaluated embeddings for captioning. CLIP-Score (Hessel et al., 2021) provides reference-free semantic alignment, while InfoMetIC (Hu et al., 2023) assesses informativeness. YOLOv8 (Talib et al., 2024) has been adopted for object detection in UAV (Zhai et al., 2023; Wang et al., 2023) and general scenarios (Rasheed and Zarkoosh, 2025).

3 Methodology

Figure 2 illustrates the VLCE pipeline. The framework works in two steps: 1) first, it creates basic captions using a ready-made VLM, and 2) it improves those captions with extra knowledge using a trained sequential model.

3.1 Baseline Caption Generation

We generate initial captions using two VLMs with complementary characteristics. YOLOv8 (Talib et al., 2024) first performs object detection on each input image I, producing an annotation set $\mathcal{A} = \{(b_i, \ell_i)\}_{i=1}^{K}$ where the bounding box coordinates $b_i \in \mathbb{R}^4$ and the class label $\ell_i \in \mathcal{L}$. These annotations are integrated into a textual prompt $P(\mathcal{A})$ describing the scene to guide the VLM.

LLaVA-7B v1.5. We use the LLaVA-7B v1.5 model (Liu et al., 2023), made of a CLIP ViT-L/14 visual encoder with a LLaMA language decoder. The encoder produces patch embeddings $V = \phi(I) = [v_1, \ldots, v_N] \in \mathbb{R}^{N \times D}$, concatenated with the prompt into sequence s. The decoder generates the caption autoregressively: $\hat{y} = \arg\max_y P(y \mid s)$. Generation uses $T = 0.2$ with a maximum output length of 4,000 tokens.

QwenVL-7B Instruct. We use the QwenVL-7B Instruct model (Bai et al., 2023) with greedy decoding and a maximum output length of 128 tokens. The vision encoder creates $V = \text{QwenVL}_{\text{vision}}(I) \in \mathbb{R}^{H \times W \times D}$. An event type e and annotations are included in a structured prompt $P(\mathcal{A}, e)$: $\hat{y} = \arg\max_y P(y \mid V, P(\mathcal{A}, e))$.

These two baselines differ substantially in output characteristics. LLaVA produces longer, more detailed descriptions, while QwenVL generates concise captions.

3.2 Knowledge Graph Vocabulary Enrichment

The vocabulary enrichment process constructs a domain-specific lexicon from the training captions augmented with external knowledge:

Step 1: Keyword extraction. We apply the RAKE algorithm to all training captions. With it we extract ranked keyword phrases that capture disaster-relevant concepts such as "debris field," "structural damage," and "emergency response".

Step 2: ConceptNet expansion. For each extracted keyword, we query the ConceptNet API, retrieving up to 10 semantically related terms per keyword. Relation filtering omits synonyms to focus on conceptual connections (e.g., "hurricane" → "wind," "flooding," "evacuation").

Step 3: WordNet synonym enrichment. Each keyword is additionally queried against WordNet to retrieve synonyms, providing lexical variety. Overlapping substrings are removed to prevent redundancy.

Step 4: Vocabulary construction. We filter out invalid English words and remove duplicates from the union of the original caption words, ConceptNet terms, and WordNet synonyms. This gives us a final vocabulary of 3,195 tokens, of which 1,566 (49%)

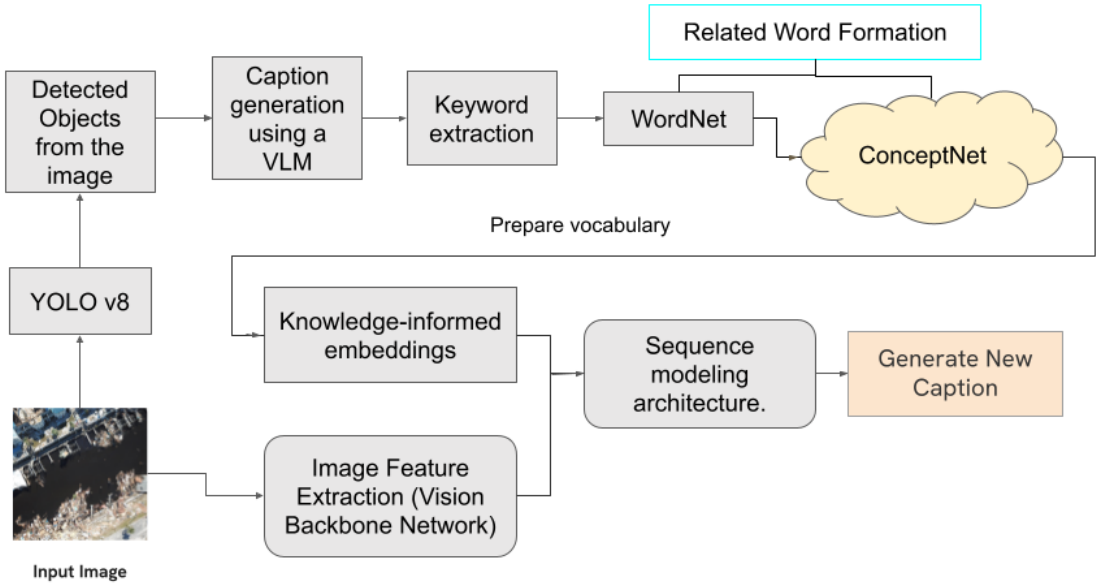


Figure 2: The VLCE pipeline. YOLOv8 generates scene-aware prompts for a baseline VLM (LLaVA or QwenVL) based on object detection in the input image. ConceptNet and WordNet queries are used to enhance keywords that were taken from training captions using RAKE, creating a vocabulary unique to the domain. The final caption is produced by training a sequential model with the enriched vocabulary and matching embeddings.

are terms that were not in the original training captions but were added to the knowledge graph. Boundary tokens (`startseq`, `endseq`) are added, and sequences are padded to max length 192.

3.3 Embedding Strategies

We evaluate two embedding strategies that differ in whether they encode structured knowledge:

With knowledge graph (KG). ConceptNet Numberbatch embeddings are 300-dimensional word vectors. They are pre-trained to encode both distributional statistics and relational knowledge from ConceptNet. Each term t is mapped to:

$$\mathbf{e}_t = \begin{cases} \text{find}(D, t) & \text{if } t \in D \\ \mathbf{e}_{\text{rand}} \sim \mathcal{U}(-\epsilon, \epsilon) & \text{otherwise} \end{cases}$$

The embedding layer is initialized with these vectors. It remains frozen during training, preserving structured semantic relationships. The final embedding matrix is $E \in \mathbb{R}^{|V| \times 300}$.

Without KG. DistilBERT embeddings (768-dimensional) provide contextual representations learned from large-scale text corpora without explicit knowledge graph structure: $e_w = \text{BERT}(w)_{[\text{CLS}]}$. This serves as an ablation to isolate the contribution of structured knowledge.

3.4 Image Encoders

We select image encoders pre-trained on data distributions matching our target domains:

ResNet50-EuroSAT. For satellite imagery (xBD), we use a ResNet50 model pre-trained on the EuroSAT land-use classification dataset[1], producing 2,048-dimensional feature vectors. Images are resized to 336×336 pixels.

ViT-UAV. For UAV imagery (RescueNet), we use a Vision Transformer pre-trained on a UAV image classification task[2], producing 768-dimensional fea-

[1] https://huggingface.co/cm93/resnet50-eurosat
[2] https://huggingface.co/SeyedAli/Remote-Sensing-UAV-image-classification

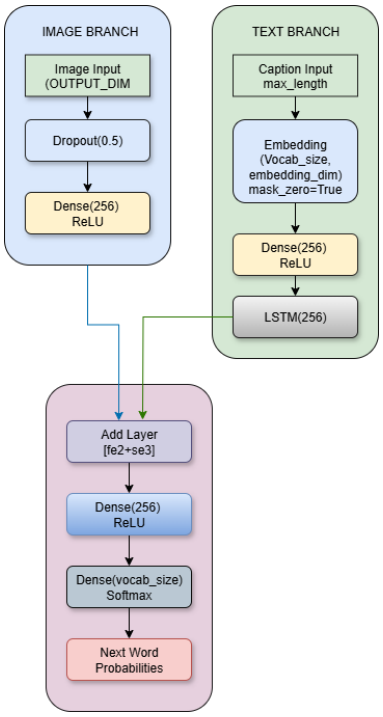


Figure 3: CNN-LSTM architecture with visual and textual branches.

ture vectors via global average pooling. Images are resized to 224×224 pixels.

3.5 Sequential Decoders

We design two decoder architectures to investigate how different sequence modeling approaches interact with knowledge-enriched vocabularies.

3.5.1 CNN-LSTM Decoder

The CNN-LSTM decoder (Figure 3) uses additive fusion of visual and textual features. Given an image feature vector $\mathbf{v} \in \mathbb{R}^{d_v}$ (where $d_v = 2048$ for ResNet50 or $d_v = 768$ for ViT-UAV) and caption tokens $w_1, \ldots, w_T$:

$$\begin{aligned} \mathbf{f}_{\text{img}} &= \text{ReLU}(\mathbf{W}_v \cdot \text{Dropout}_{0.5}(\mathbf{v})) \in \mathbb{R}^{256} \\ \mathbf{f}_{\text{txt}} &= \text{LSTM}_{256}(\text{Dropout}_{0.5}(\mathbf{E}[w_1, \ldots, w_T])) \in \mathbb{R}^{256} \\ \mathbf{h} &= \text{ReLU}(\mathbf{W}_h \cdot (\mathbf{f}_{\text{img}} + \mathbf{f}_{\text{txt}})) \in \mathbb{R}^{256} \\ P(w_{t+1} \mid w_{1:t}, \mathbf{v}) &= \text{softmax}(\mathbf{W}_o \cdot \mathbf{h}) \end{aligned} \tag{1}$$

where $\mathbf{E}$ is the embedding matrix (initialized from ConceptNet Numberbatch or DistilBERT and frozen), and $\mathbf{W}_v, \mathbf{W}_h, \mathbf{W}_o$ are learned projection matrices.

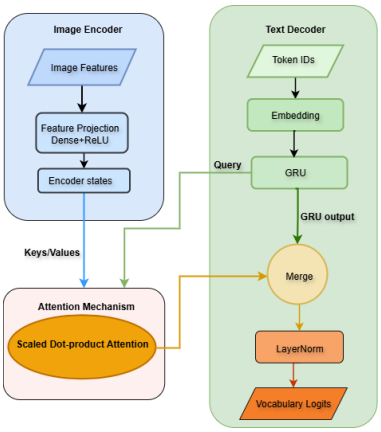


Figure 4: Hierarchical cross-modal Transformer architecture.

3.5.2 Hierarchical Cross-Modal Transformer Decoder

The Transformer decoder (Figure 4) introduces a hierarchical visual encoding scheme that captures multi-scale spatial information, followed by cross-modal attention for visually grounded text generation.

Multi-scale visual encoding. The image feature vector $\mathbf{v} \in \mathbb{R}^{d_v}$ is projected into three complementary representations:

$$\begin{aligned} \mathbf{f}_{\text{global}} &= \text{ReLU}(\mathbf{W}_g \mathbf{v} + \mathbf{b}_g) \in \mathbb{R}^{d_{\text{model}}} \\ \mathbf{F}_{\text{regional}} &= \text{Dense}_{d_{\text{model}}}\left(\text{Reshape}_{4 \times d_{\text{model}}/4}(\text{ReLU}(\mathbf{W}_r \mathbf{v} + \mathbf{b}_r))\right) \in \mathbb{R}^{4 \times d_{\text{model}}} \\ \mathbf{F}_{\text{local}} &= \text{Dense}_{d_{\text{model}}}\left(\text{Reshape}_{12 \times d_{\text{model}}/12}(\text{ReLU}(\mathbf{W}_l \mathbf{v} + \mathbf{b}_l))\right) \in \mathbb{R}^{12 \times d_{\text{model}}} \end{aligned} \tag{2}$$

These are concatenated and normalized to form the visual context:

$$\mathbf{F}_v = \text{LayerNorm}([\mathbf{f}_{\text{global}}; \mathbf{F}_{\text{regional}}; \mathbf{F}_{\text{local}}]) \in \mathbb{R}^{17 \times d_{\text{model}}} \tag{3}$$

Cross-modal decoding. Input tokens are embedded with frozen pre-trained matrix $\mathbf{E}$ and sinusoidal positional encodings: $\mathbf{H}^{(0)} = \mathbf{E}[\mathbf{w}_{1:T}] + \text{PE}(\mathbf{w}_{1:T})$. Each of the L decoder layers applies causal self-attention over caption tokens, followed by cross-attention over the visual context:

$$\begin{aligned} \mathbf{A}_{\text{self}} &= \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}_{\text{causal}}) \\ \mathbf{A}_{\text{cross}} &= \text{MultiHead}(\mathbf{H}, \mathbf{F}_v, \mathbf{F}_v) \end{aligned} \tag{4}$$

where $\mathbf{H}$ is the output of the self-attention sublayer and $\mathbf{M}_{\text{causal}}$ enforces the autoregressive property.

Output generation. Global visual semantics are fused with textual representations:

$$\begin{aligned} \mathbf{c}_{\text{visual}} &= \text{GlobalAvgPool}(\mathbf{F}_v) \\ \hat{\mathbf{y}}_t &= \text{Dense}_V\left(\text{LayerNorm}\left([\mathbf{h}_t^{(L)}; \mathbf{c}_{\text{visual}}]\right)\right) \end{aligned} \tag{5}$$

3.6 Training

All models use $d_{\text{model}} = d_{\text{emb}} = 300$, $L = 2$ Transformer layers with $h = 6$ attention heads, and are trained with masked sparse categorical cross-entropy loss:

$$\mathcal{L} = -\frac{1}{\sum_{t=1}^{T} m_t} \sum_{t=1}^{T} m_t \log P(w_t \mid \mathbf{w}_{<t}, \mathbf{F}_v; \theta) \tag{6}$$

where $m_t \in \{0, 1\}$ masks padding positions. Batch size is 32, maximum sequence length 192 tokens. Captions are bounded by `startseq` and `endseq` tokens. At inference, generation proceeds autoregressively until the stop token or maximum length is reached. The complete pipeline is formalized in Algorithm 1.

Algorithm 1 VLCE: Knowledge Graph-Enhanced Caption Refinement

Require: Disaster image dataset $\mathcal{D} = \{I_1, \ldots, I_n\}$
Ensure: Enhanced captions $\mathcal{C}_{\text{Enhanced}}$

Stage 1: Baseline Caption Generation
for each image $I_i \in \mathcal{D}$ **do**
  $O_i \leftarrow \text{YOLOv8}(I_i)$; $P_i \leftarrow \text{ConstructPrompt}(I_i, O_i)$
  $C_i^{\text{VLM}} \leftarrow \text{VLM}(I_i, P_i)$
end for

Stage 2: Knowledge Preparation
$\mathcal{V}_{\text{Enriched}} \leftarrow \text{Vocab}(\{C_i^{\text{VLM}}\}) \cup \text{RAKE+ConceptNet+WordNet}$
$E \leftarrow \text{BuildEmbeddings}(\mathcal{V}_{\text{Enriched}})$

Stage 3: Training
$\mathcal{F}_{\text{Img}} \leftarrow E_{\text{Visual}}(\mathcal{D})$; Train $\mathcal{N}_\theta$ on $(\mathcal{F}_{\text{Img}}, \mathcal{V}_{\text{Enriched}}, E)$

Stage 4: Enhanced Caption Generation
for each image I_i **do**
  $C_i^{\text{Enhanced}} \leftarrow \text{Decode}(I_i, \mathcal{N}_\theta)$
end for
return $\mathcal{C}_{\text{Enhanced}}$

4 Experimental Setup

4.1 Datasets

xBD (Gupta et al., 2019).A satellite imagery dataset for building damage assessment containing 12,738 images. We use 6,369 post-disaster images, merging "major-damage" and "minor-damage" into a single "damaged" class to yield three labels: no-damage, damaged, and destroyed. Split: 80/20 (5,095 train / 1,274 test).RescueNet (Rahnemoonfar et al., 2023). A UAV imagery dataset captured after Hurricane Michael, containing 4,494 images at $3{,}000 \times 4{,}000$ pixel resolution across 12 fine-grained damage categories. Split: 80/20.

4.2 Evaluation Metrics

We evaluate using two complementary metrics: CLIPScore (Hessel et al., 2021), which computes the cosine similarity between CLIP embeddings (`openai/clip-vit-base-patch32`) of the image and generated caption. For each test image, we compare the CLIPScore of the VLCE-generated caption against the baseline VLM caption and note the percentage of images where VLCE scores higher. Another one is InfoMetIC (Hu et al., 2023), a composite metric: $\alpha \cdot \text{Informativeness} + \beta \cdot \text{Relevance} + \gamma \cdot \text{Precision}$, where Informativeness is computed as the average negative log-probability of caption words conditioned on the image using a vision-language model. InfoMetIC penalizes generic descriptions and rewards information-rich captions.

4.3 Configurations

We associate each decoder architecture with a designated baseline VLM: CNN-LSTM with LLaVA and Transformer with QwenVL. Each pairing is assessed with and without knowledge graph enhancement across both datasets (RescueNet/ViT-UAV and xBD/ResNet50-EuroSAT).

5 Results

5.1 Main Results

Table 1 presents the percentage of test images where VLCE captions score higher than the corresponding baseline VLM caption. Values above 50% indicate VLCE outperforms the baseline for the majority of images.

5.2 Key Findings

Knowledge graph integration is critical for UAV imagery. On RescueNet, removal of knowledge graph (KG) enrichment causes near-complete performance collapse for both architectures (Figure 8). The CNN-LSTM without KG achieves only 0.56% CLIPScore preference and 1.22% InfoMetIC preference. The Transformer without KG drops to 0.22% and 0.08%, respectively. The 12-class damage taxonomy of RescueNet requires exact terminology that

Table 1: Percentage of test images where VLCE captions score higher than the baseline VLM caption. CNN-LSTM models are compared against LLaVA; Transformer models against QwenVL. Bold indicates VLCE outperforms the baseline on the majority of images.

				CLIPScore (%)		InfoMetIC (%)	
Decoder	**Baseline**	**Dataset**	**KG**	VLCE	Baseline	VLCE	Baseline
CNN-LSTM	LLaVA	RescueNet	Yes	**52.95**	47.05	**54.51**	45.49
		RescueNet	No	0.56	99.44	1.22	98.78
		xBD	Yes	**51.10**	48.90	**66.56**	33.44
		xBD	No	**55.34**	44.66	**66.41**	33.59
Transformer	QwenVL	RescueNet	Yes	**73.64**	26.36	**95.33**	4.67
		RescueNet	No	0.22	99.78	0.08	99.92
		xBD	Yes	**60.60**	39.40	**69.86**	30.14
		xBD	No	22.14	77.86	18.76	81.24

Table 2: Relevant object counts across model configurations.

Exp.	**Configuration**	**VLCE**	**LLaVA**	**QwenVL**	**Best**
	ViT-UAV / RescueNet				
1	With Knowledge Graph	**272**	185	201	VLCE
2	Without Knowledge Graph	**272**	178	195	VLCE
	ResNet-EuroSAT / xBD				
3	With Knowledge Graph	**640**	445	463	VLCE
4	Without Knowledge Graph	**640**	431	448	VLCE
	Summary				
Avg. improvement over LLaVA		**+45.2%**	–	–	–
Avg. improvement over QwenVL		**+38.7%**	–	–	–
Experiments won		**4/4**	0/4	0/4	–

the model cannot learn from limited training captions alone.

The Transformer with KG achieves the strongest results. The best configuration is the Transformer decoder with ViT-UAV encoder and KG enrichment on RescueNet, achieving 73.64% CLIPScore and 95.33% InfoMetIC preference over QwenVL (Figure 5, Figure 6). The gap between metrics indicates VLCE captions are substantially more *informative* than the QwenVL baseline, even in cases where visual alignment scores are competitive.

Satellite imagery is more forgiving. On xBD, CNN-LSTM without KG still achieves 55.34% CLIPScore and 66.41% InfoMetIC preference. The lower label complexity (3 vs. 12 classes) and the ResNet50-EuroSAT encoder's satellite-specific pre-training partially compensate for the absence of knowledge-enriched vocabulary.

Architecture-specific patterns. The CNN-LSTM shows binary behavior on RescueNet: it works with KG or fails without. The Transformer exhibits more graceful degradation on xBD without KG. It achieves a 22.14% CLIPScore, suggesting cross-attention can partially compensate for missing vocabulary knowledge on simpler scenes.

5.3 Quantitative Analysis

Figures 5–8 aggregate the quantitative findings. Two observations are noteworthy. The KG ablation (Figure 7) demonstrates that the Transformer architecture derives much bigger advantages from KG on RescueNet (+73.4 pp CLIPScore, +95.3 pp InfoMetIC) compared to xBD (+38.5 pp, +51.1 pp), indicating the heightened language requirements of UAV images. Secondly, the CNN-LSTM on xBD without knowledge graphs is the sole configuration in which the removal of knowledge graphs marginally

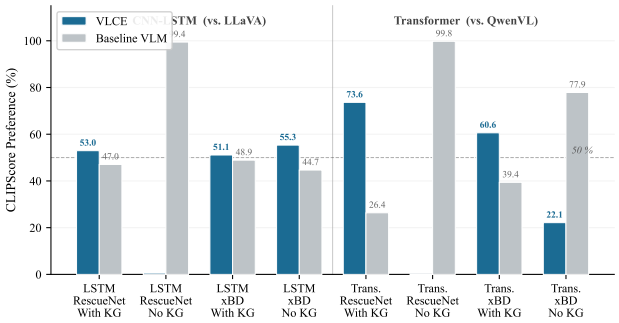


Figure 5: CLIPScore preference across all configurations. The dashed line marks 50% (parity). Without KG, VLCE preference collapses on RescueNet.

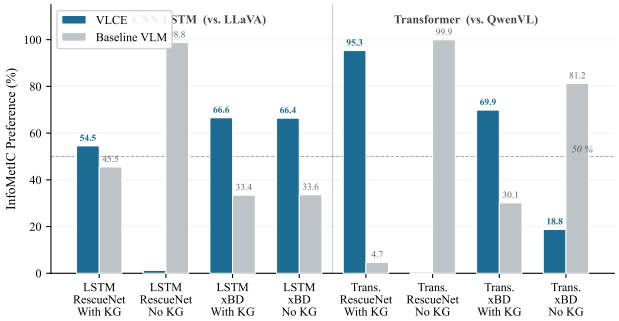


Figure 6: InfoMetIC preference across all configurations. The Transformer with KG on RescueNet achieves 95.33% preference over QwenVL.

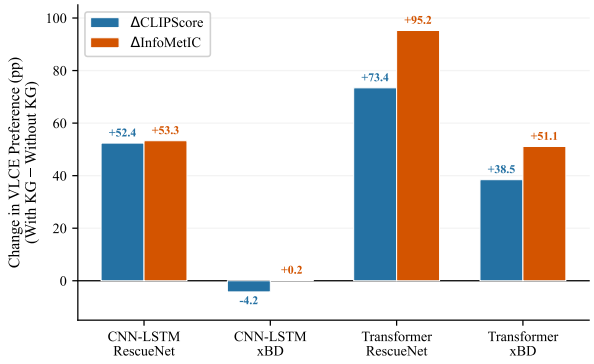


Figure 7: Effect of KG integration measured as the change in VLCE preference (pp). The Transformer on RescueNet gains +73.4 pp CLIPScore and +95.3 pp InfoMetIC.

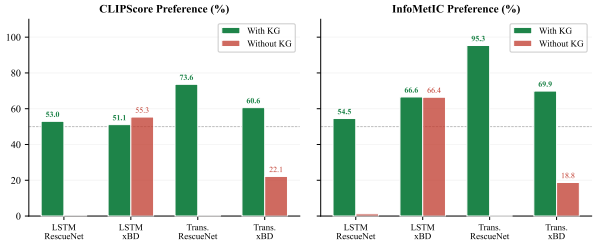


Figure 8: With vs. without KG. Green bars consistently exceed 50%, while red bars collapse to near-zero on RescueNet.

enhances the CLIPScore by 4.2 percentage points, indicating that for satellite imagery with less complex damage categories, direct visual-linguistic mapping can be competitive. Appendix A.1 contains the distributions of per-image scores.

5.4 Noun-Based Object Detection Analysis

Table 2 reports the total count of unique relevant nouns detected across all test captions to assess whether VLCE captures more disaster-relevant objects.

VLCE achieves full coverage of the reference object set (272 UAV objects, 640 satellite objects) compared to 55–65% by baseline VLMs, confirming the knowledge-enriched vocabulary enables broader and more precise object references.2

6 Discussion

Why knowledge graphs matter more for UAV imagery. The significant performance gap between with-KG and without-KG configurations on RescueNet vs. xBD shows the main differences between the two imaging methods. UAV imagery captures scenes from oblique angles at high spatial resolution. It reveals fine-grained damage details such as collapsed walls, scattered debris, and standing water that require precise vocabulary to describe. Satellite imagery, observed from a direct overhead perspective at lower resolution, shows more uniform visual patterns where more abstract descriptions suffice. RescueNet's 12-class damage taxonomy also necessitates specific terminology, which can only be provided by the knowledge-enriched vocabulary.

Complementarity of CLIPScore and InfoMetIC. Configurations with a high InfoMetIC but a moderate CLIPScore, such as the Transformer with KG on RescueNet (95.33% vs. 73.64%), show that using a knowledge graph helps the model include important details that aren't directly visible in the image. CLIPScore, which assesses image-text alignment, may underestimate descriptions that include domain knowledge regarding disaster processes, whereas InfoMetIC values such informational depth.

Practical implications. For operational disaster response, captions must go beyond visible elements. InfoMetIC-favored captions, which include

domain-appropriate terminology and contextual reasoning, are more actionable, as demonstrated by the Transformer+KG configuration's strong performance. This suggests knowledge graphs bridge the gap between visual observation and situational understanding.

7 Conclusion

We introduced VLCE, a framework for generating knowledge-enriched captions for post-disaster imagery. By integrating external semantic knowledge from ConceptNet and WordNet into the caption generation pipeline, VLCE addresses the vocabulary and domain knowledge limitations of general-purpose VLMs when applied to disaster assessment. Our experiments on the xBD and RescueNet benchmarks demonstrate that knowledge graph enrichment is critical for UAV imagery captioning, where it transforms near-zero performance into strong preference over baseline VLMs. The Transformer decoder with knowledge graph integration achieves the best overall results, with 95.33% InfoMetIC preference on RescueNet. Qualitative analysis substantiates that knowledge-enhanced captions preserve factual accuracy and contextually relevant terminology, whereas captions produced without knowledge graph assistance display hallucinations, redundancy, and semantic disparity.

Acknowledgement: We used ChatGPT 5 to assist in refining the language, clarity, and formatting of the LaTeX manuscript. All technical content, analysis, and conclusions are the authors' own.

References

O. Abbas and J. Dang. 2023. Using image captioning for automatic post-disaster damage detection. *Intelligence, Informatics and Infrastructure*, 4(2):66–74.

I. Alisjahbana and 1 others. 2024. DeepDamageNet: A two-step deep-learning model for multi-disaster building damage segmentation. *Preprint*, arXiv:2405.04800.

J. Bai and 1 others. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

L. Chen and K. Li. 2025. Multi-modal graph aggregation transformer for image captioning. *Neural Networks*, 181:106813.

P.-J. Chun, T. Yamane, and Y. Maemura. 2022. A deep learning-based image captioning method for bridge damage. *Computer-Aided Civil and Infrastructure Engineering*, 37(11):1387–1401.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

S. Elbedwehy and 1 others. 2023. Enhanced image captioning using features concatenation and efficient pre-trained word embedding. *Computer Systems Science and Engineering*, 46:3637–3652.

Ritwik Gupta and 1 others. 2019. xBD: A dataset for assessing building damage from satellite imagery. *Preprint*, arXiv:1911.09296.

J. Hessel and 1 others. 2021. CLIPScore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

A. Hu and 1 others. 2023. InfoMetIC: An informative metric for reference-free image caption evaluation. *Preprint*, arXiv:2305.06002.

S. Kota and 1 others. 2022. Multimodal combination of text and image tweets for disaster response assessment. In *D2R2*.

H. Liu and 1 others. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Rebecca Mason and Eugene Charniak. 2014. Domain-specific image captioning. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 28–36. Association for Computational Linguistics.

M. Rahnemoonfar, T. Chowdhury, and R. Murphy. 2023. RescueNet image classification dataset. Figshare Dataset.

A. F. Rasheed and M. Zarkoosh. 2025. Optimized YOLOv8 for multi-scale object detection. *Journal of Real-Time Image Processing*, 22(1):6.

C. Schön, S. Siebert, and F. Stolzenburg. 2019. Using ConceptNet to teach common sense to an automated theorem prover. *Electronic Proceedings in Theoretical Computer Science*, 311:19–24.

M. Talib, A. H. Y. Al-Noori, and J. Suad. 2024. YOLOv8-CAB: Improved YOLOv8 for real-time object detection. *Karbala International Journal of Modern Science*, 10(1):5.

J. Tang, P. Li, and M. Jiang. 2022. Image caption generation based on knowledge graph guidance and self-attention. Preprint.

M. Toshevska, F. Stojanovska, and J. Kalajdjieski. 2020. Comparative analysis of word embeddings for capturing word similarities. *Preprint*, arXiv:2005.03812.

G. Wang and 1 others. 2023. UAV-YOLOv8: A small-object-detection model for UAV aerial photography. *Sensors*, 23(16):7190.

X. Wang and 1 others. 2024. Knowledge enhancement strategies for remote sensing image captioning using CLIP and LLMs. In *Proceedings of MIDA*, pages 313–318.

X. Zhai and 1 others. 2023. YOLO-Drone: An optimized YOLOv8 for tiny UAV object detection. *Electronics*, 12(17):3664.

Wentian Zhao and Xinxiao Wu. 2024. Boosting entity-aware image captioning with multi-modal knowledge graph. *IEEE Transactions on Multimedia*, 26:2659–2670.

Y. Zhou, Y. Sun, and V. Honavar. 2019. Improving image captioning by leveraging knowledge graphs. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 539–548.

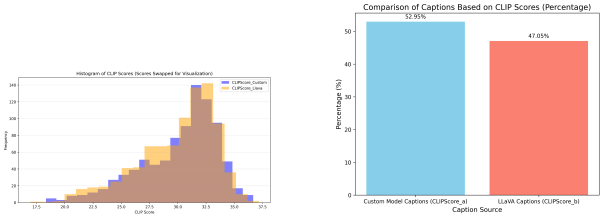
Z. Zou and 1 others. 2021. Disaster image classification by fusing multimodal social media data. *ISPRS International Journal of Geo-Information*, 10(10):636.

A Appendix

A.1 Per-Image Score Distributions

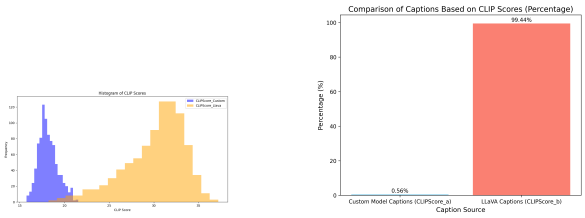
This appendix provides per-image CLIPScore and InfoMetIC score distributions for all configurations. These histograms and bar charts complement the summary statistics in Table 1.

A.1.1 CLIPScore Distributions



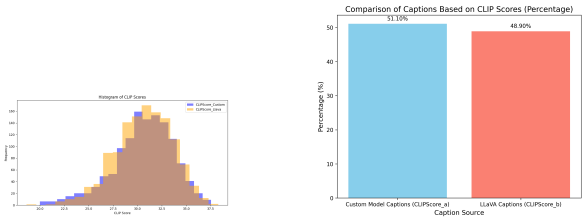
(a) Score distribution (b) Model comparison

Figure 9: CLIPScore: RescueNet + LLaVA **with** KG.



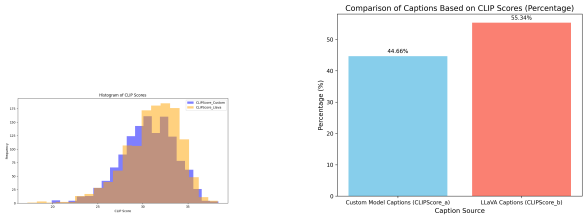
(a) Score distribution (b) Model comparison

Figure 10: CLIPScore: RescueNet + LLaVA **without** KG.



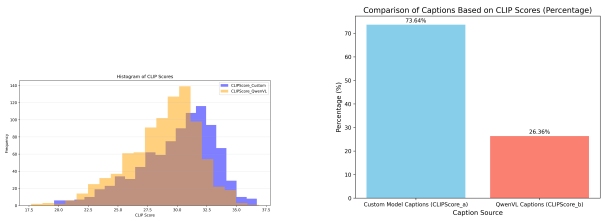
(a) Score distribution (b) Model comparison

Figure 11: CLIPScore: xBD + LLaVA **with** KG.



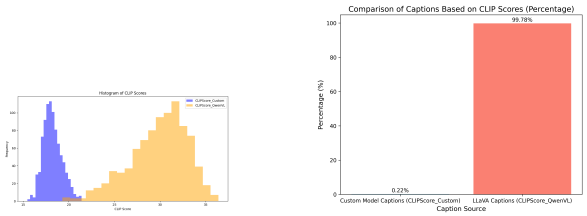
(a) Score distribution (b) Model comparison

Figure 12: CLIPScore: xBD + LLaVA **without** KG.



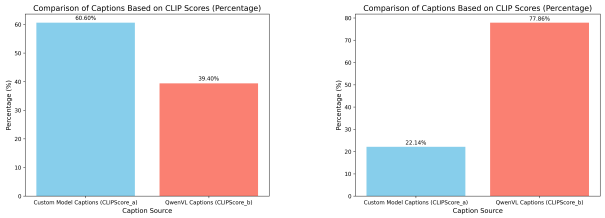
(a) Score distribution (b) Model comparison

Figure 13: CLIPScore: RescueNet + QwenVL **with** KG.



(a) Score distribution (b) Model comparison

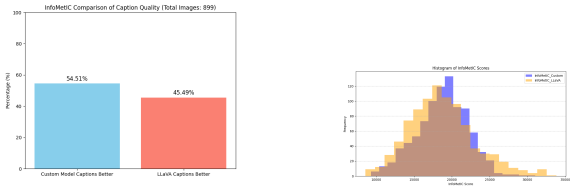
Figure 14: CLIPScore: RescueNet + QwenVL **without** KG.



(a) With KG (b) Without KG

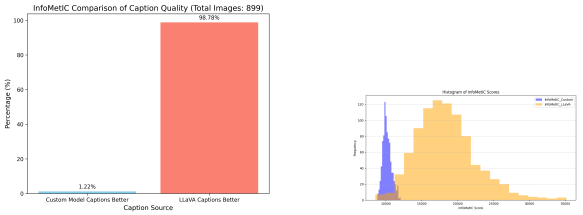
Figure 15: CLIPScore: xBD + QwenVL, with vs. without KG.

A.1.2 InfoMetIC Distributions



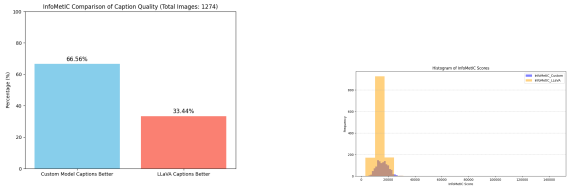
(a) Model comparison (b) Score distribution

Figure 16: InfoMetIC: RescueNet + LLaVA **with** KG.



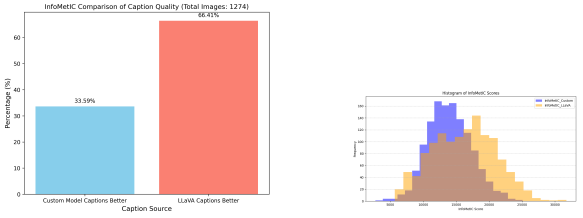
(a) Model comparison (b) Score distribution

Figure 17: InfoMetIC: RescueNet + LLaVA **without** KG.



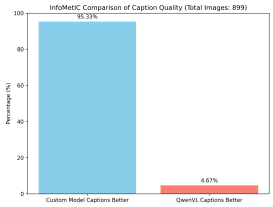
(a) Model comparison (b) Score distribution

Figure 18: InfoMetIC: xBD + LLaVA **with** KG.

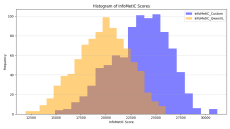


(a) Model comparison (b) Score distribution

Figure 19: InfoMetIC: xBD + LLaVA **without** KG.

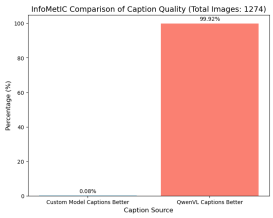


(a) Model comparison

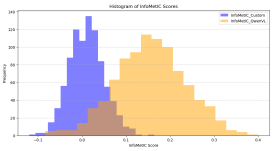


(b) Score distribution

Figure 20: InfoMetIC: RescueNet + QwenVL **with** KG.

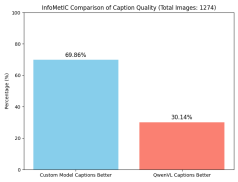


(a) Model comparison

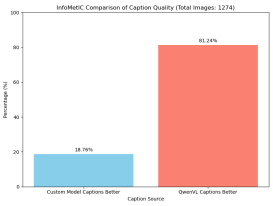


(b) Score distribution

Figure 21: InfoMetIC: RescueNet + QwenVL **without** KG.



(a) With KG



(b) Without KG

Figure 22: InfoMetIC: xBD + QwenVL, with vs. without KG.

A.2 Qualitative Case Studies

We examine representative examples across dataset-model combinations to illustrate how KG integration affects caption quality.

A.2.1 RescueNet + LLaVA

KG-Enhanced Caption [23]

Caption: the image depicts the aftermath of hurricane michael which caused significant damage to the infrastructure and environment in the affected area **the most noticeable feature is the flooded area which has submerged the debris from damaged structures can be seen scattered throughout the scene** with some debris piled up in the water the presence of multiple vehicles including cars and trucks indicates that the area is being cleared and assessed for recovery efforts the overall scene conveys sense of destruction and recovery efforts needed to restore the affected area to its previous state.

The KG-enhanced caption coherently identifies flooding as the primary damage type, notes submerged debris, and infers recovery operations from visible vehicles.

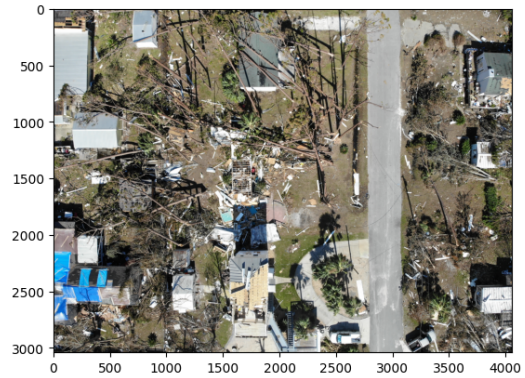


Figure 23: RescueNet with KG-enhanced LLaVA output.

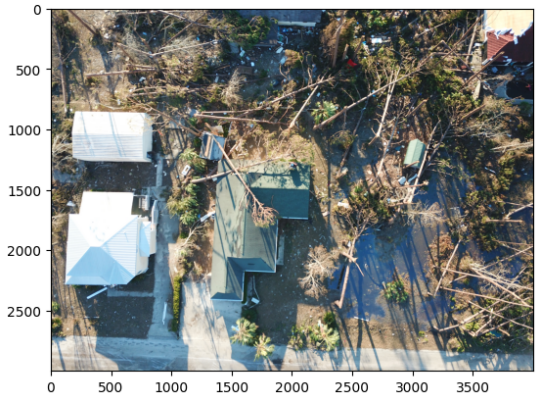


Figure 24: RescueNet without KG, LLaVA output.

Baseline Caption (No KG) [24]

Caption: **the the** image shows an aerial view of a residential severely impacted by a natural disaster fallen trees lie across rooftops, **stone** and streets directly visible structural damage can be seen while the surrounding area is littered with broken branches and debris the stark contrast between intact rooftops and **widespreadvegetation** damage impacted **by a natural disaster fallen trees lie rooftops, stone**

Without KG: word duplication ("the the"), nonsensical terms ("stone"), concatenated words ("widespreadvegetation"), and verbatim phrase repetition.

A.2.2 xBD + LLaVA

KG-Enhanced Caption [25]

Caption: the image shows satellite view of forest with clear path through the trees the path is surrounded by trees and there are several trees in the area the pathway appears to be dirt road and it is surrounded by trees the **image also shows few buildings in the area which seem to have escaped the aftermath of the disaster** the image captures the aftermath of disaster with undamaged areas or buildings still standing amidst the destruction

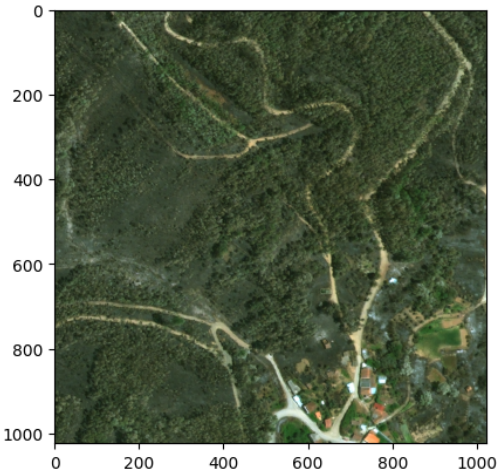


Figure 25: xBD with KG-enhanced LLaVA output.

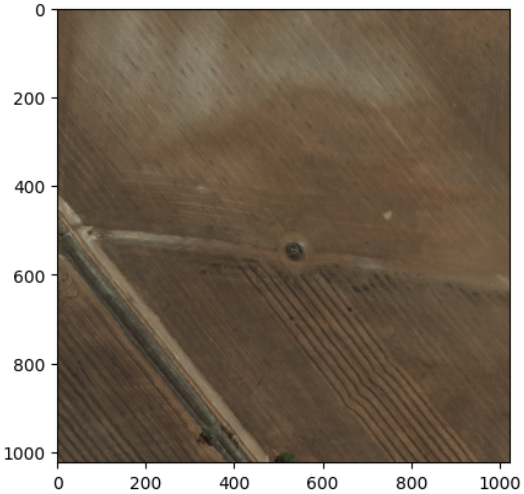


Figure 26: xBD without KG, LLaVA output.

Baseline Caption (No KG) [26]

Caption: the image shows satellite view of field with large crack in the middle likely caused by an earthquake **the crack extends from the left side of the image to the right dividing the field into two distinct sections additionally there are two people in the field one near the center** and the other towards the right side the presence of these holes suggests that the area has been affected by the disaster and**the animals deaths may have been temporarily closed due to the damage**

The baseline fabricates elements (people in the field, animal deaths) and produces semantically incoherent phrases—a dangerous failure mode for disaster assessment.

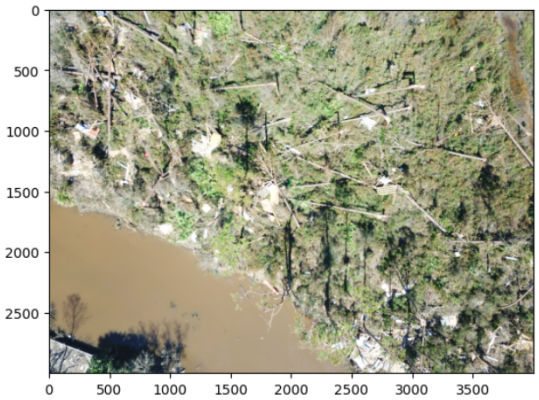


Figure 27: RescueNet with KG-enhanced QwenVL output.

A.3 RescueNet + QwenVL

KG-Enhanced Caption [27]

Caption: the satellite image depicts an area that has been significantly impacted by hurricane the **image shows dense concentration of trees that appear to be damaged or fallen indicating the powerful force of the storm the trees are scattered** throughout the area with some lying on the ground and others leaning precariously **the extent of the destruction is widespread affecting both the natural landscape** and built environments the fallen trees and debris suggest that **the hurricane had strong impact on the vegetation potentially causing widespread damage to the local ecosystem** the presence of debris and fallen trees can obstruct roads and paths making it difficult for emergency services and residents

Figure 28: RescueNet without KG, QwenVL output.

Baseline Caption (No KG) [28]

Caption: the satellite photo depicts an area that has been significantly impacted by **a food area and causing harm to five people** of the destruction is evident from the scattered debris and damaged structures here is detailed description of the visible damage and its implications damaged structures there are two large piles of debris scattered across the area these piles consist of various materials such as trees, wood metal and possibly concrete indicating that multiple buildings have been severely damaged or destroyed the debris

The baseline generates semantic inaccuracies ("food area and causing harm to five people"), fabricated casualty figures, and fragmented sentences.

A.4 xBD + QwenVL

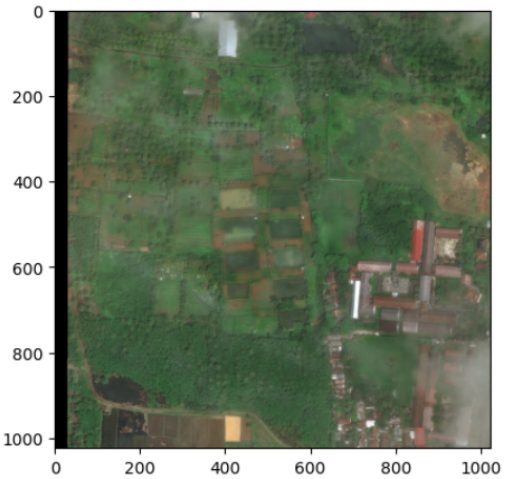


Figure 29: xBD with KG-enhanced QwenVL output.

KG-Enhanced Caption [29]

Caption: The image shows **satellite view of town with mix of undamaged and damaged buildings there are undamaged buildings and damaged buildings visible in the scene** the presence of both undamaged and damaged structures highlights the impact of the disaster on the affected area

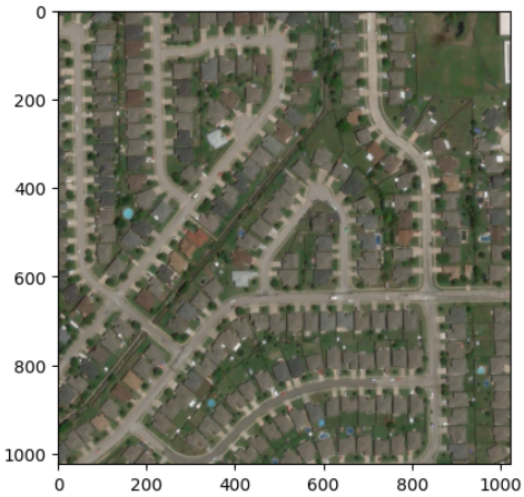


Figure 30: xBD without KG, QwenVL output.

Baseline Caption (No KG) [30]

Caption: the satellite image depicts an area that has been significantly impacted by tornado the visible damage is extensive with numerous buildings and structures **showing signs of destruction showing signs of destruction** here is detailed description of the visible damage and its implications damage overview damaged areas buildings these areas show visible signs of damage such as showing signs of destruction **collapsed trees broken** and debris scattered around the damage is concentrated in specific areas indicating that the tornado had localized impact undamaged **areasbuildings** these areas appear to be relatively intact with no visible signs of damage from the tornado

A.5 Summary of Failure Modes

Across all configurations, removing KG consistently produces four categories of errors: (1) **hallucinations**- fabricated casualty counts and fictitious elements; (2) **redundancy**- phrases repeated three or more times; (3) **formatting errors**- concatenated words and missing spacing; and (4) **semantic incoherence**- fragmented sentences. KG-enhanced models eliminate these failure modes across all architectures, VLMs, and datasets.