Evaluating Visual Language Models for Multimodal Crisis Classification

Christine N. Muthee
Carnegie Mellon University
Pittsburgh, PA, USA
cmuthee@andrew.cmu.edu

Keriane L. Nzabampema
Carnegie Mellon University Africa
Kigali, Rwanda
lnzabampema@andrew.cmu.edu

***Abstract*—The increasing frequency of natural disasters highlights the urgent need for rapid and accurate situational awareness to support the humanitarian response. Social media platforms like Twitter offer multimodal real-time content, but extracting actionable information from noisy image-text pairs remains a significant challenge. Existing approaches, such as CrisisKAN, rely on external knowledge sources such as Wikipedia to improve performance, but these sources are often unavailable during the initial disaster phases. In this study, we investigate whether Visual Language Models (VLMs), specifically through automated image captioning, can serve as a real-time alternative to Wikipedia-based knowledge infusion in multimodal crisis classification. Using the CrisisMMD dataset, we evaluated three model variants: (1) a baseline CrisisKAN model without external knowledge, (2) a caption-enhanced multimodal model trained in both single-task and multi-task settings, and (3) a fine-tuned Contrastive Language–Image Pretraining (CLIP) model used as a lightweight classification alternative. Our results show that knowledge-infused models outperform the non-infused baseline in informativeness detection (Task 1), while tasks involving fine-grained damage or humanitarian categories (Tasks 2 and 3) remain challenging due to class imbalance and limited data. Although CLIP offers a faster, less resource-intensive option, it lags in accuracy compared to models with explicit fusion. These findings support the feasibility of using VLMs for scalable, real-time disaster response systems, especially severity ranking tasks. Although they do not yet match the accuracy of CrisisKAN, the models offer a promising foundation for future crisis classification systems.**

I. Introduction

Natural disasters continue to pose significant threats to human life, infrastructure and social stability,causing an estimated 40,000 to 50,000 deaths annually worldwide [11] In the aftermath of such events, humanitarian agents must make rapid decisions under pressure, often with limited resources and time. Social media platforms such as Twitter have emerged as powerful tools for real-time situational awareness, offering a stream of posts that include both textual descriptions and visual evidence of damage [1]. However, the unstructured, high-volume nature of this data introduces major challenges. Humanitarian responders face three key issues: (1) distinguishing informative content from noise, (2) understanding the severity of damage, and (3) categorizing the type of incident, all while operating under extreme time constraints.

Traditional crisis response models have relied mainly on text-based classification systems such as the Artificial Intelligence for Disaster Response (AIDR) system [2] which processes tweets to determine their relevance. Although effective in identifying textual information related to the crisis, text-only models struggle to capture visual context which limits their ability to assess the severity of damage and provide a comprehensive situational overview [2]. Image-based models, on the other hand, have been used for disaster damage assessment but lack the ability to incorporate textual context, making it difficult to determine the full scope of an event [1]. Recent research has explored multimodal learning approaches that combine textual and visual data to improve crisis classification. The CrisisMMD dataset, for example, introduced a structured Multimodal dataset for disaster classification [1], while later work such as CrisisKAN attempt to address this by fusing image and text inputs with additional context drawn from external sources like Wikipedia [3]. While effective, these systems are not designed for real-time deployment, as they depend on slow, post-disaster Wikipedia updates and require significant computational resources to retrieve and integrate this knowledge.

In this study, we explore whether Visual Language Models (VLMs), specifically those capable of generating image captions, can serve as a lightweight, real-time alternative to Wikipedia-based knowledge infusion. We evaluated three modeling strategies using the CrisisMMD dataset: (1) a replication of CrisisKAN without knowledge infusion, (2) a caption-enhanced multimodal model, and (3) a fine-tuned CLIP model serving as a simple, unified encoder for text-image pairs.

This paper presents an in-depth comparison of these models, focusing on their performance across three crisis classification tasks: informativeness filtering, humanitarian categorization, and damage severity assessment. Our work contributes to the growing body of research on practical real-time AI systems for disaster response by assessing the feasibility, accuracy, and trade-offs of VLM-based alternatives.

The next section reviews key developments in disaster response AI, from unimodal classifiers to recent multimodal fusion approaches.

II. Related Work/ Literature Review

Artificial Intelligence has significantly shaped the landscape of disaster response by enabling rapid and automated analysis of social media data. Early work in this domain was primarily text-centric. For example, Balasubramanian et al. [4] developed a machine learning framework that classified tweets as actionable or nonactionable during the COVID-19 pandemic, improving healthcare logistics. However, this framework lacked the ability to incorporate image data, thereby limiting the situational context available to responders.

Text-only systems such as AIDR [2] further showcased the power of machine learning to automate crisis classification by ingesting and labeling Twitter data. Although AIDR achieved an accuracy of 80%, its dependency on crowd-sourced human annotation introduced latency and subjectivity, affecting real-time scalability. These models also exhibited poor performance on tasks that require spatial awareness or damage recognition due to their reliance on inputs from natural languages alone [2, 7].

To mitigate these limitations, researchers began exploring image-based classification systems. Deep learning approaches using CNNs and segmentation models such as ResNet, UNet, and PSPNet were deployed to assess damage levels from satellite imagery [5, 9]. Although effective in identifying physical destruction, these systems lacked semantic depth, since they ignored concurrent textual reports that often conveyed contextual information crucial for relief prioritization.

This led to the evolution of multimodal approaches that integrate both text and image characteristics to enhance situational awareness. Alam et al. [1] introduced CrisisMMD, the first large-scale human-annotated multimodal dataset designed for the classification of social media crisis. Despite its foundational value, initial models using CrisisMMD relied heavily on early fusion (e.g., feature concatenation), missing the opportunity for deeper cross-modal understanding.

Subsequent work by Abavisani et al. [6] proposed the use of cross-attention mechanisms to selectively integrate relevant image and text components, showing substantial improvements over concatenation-based fusion. Their graph-based augmentation technique further allowed the generation of synthetic multimodal pairs, improving generalization. Nevertheless, their framework still lacked localization capabilities and faced potential overfitting due to over-reliance on dominant modalities.

CrisisKAN [3] introduced Guided Cross-Attention Fusion, a sophisticated architecture to align image and text representations. Its integration of Wikipedia-based knowledge improved contextual depth, and its explainability module allowed more interpretable predictions. However, the practical limitations of this approach, notably the delay in Wikipedia updates and the computational cost of entity linking, render it unsuitable for real-time deployment scenarios.

Our approach builds upon CrisisKAN by replacing the static Wikipedia-based infusion with a dynamic Visual Language Model (VLM). Using phi-4, a transformer-based VLM, we generate captions for each image and concatenate them with the original tweet, thus enriching textual inputs with image semantics. Unlike works that entirely replace traditional encoders with VLM, we integrate this additional knowledge alongside CrisisKAN's original image (DenseNet) and text (ELECTRA) encoders to preserve and improve modality-specific features.

Furthermore, we test CLIP-based classifiers as a lightweight alternative that removes the need for explicit fusion mechanisms. This unifies text and image embeddings into a single space, offering improved efficiency and modularity at the cost of some fine-grained performance.

Our contributions address a fundamental research gap: enabling knowledge-rich real-time crisis classification without dependency on external databases like Wikipedia. By analyzing multiple architectures and training regimes (single-task, multi-task, frozen/fine-tuned), we demonstrate that multimodal learning can be adapted to real-world operational constraints while still preserving classification accuracy.

III. Problem Formulation

Natural disaster response depends heavily on the ability to quickly identify, categorize, and prioritize critical information from noisy real-time data sources, such as Twitter. However, existing AI systems for disaster classification face persistent limitations. Text-only models lack visual grounding and often miss contextual cues that are essential to understand the scale or type of crisis. Image-only models, on the other hand, cannot interpret scene-specific textual details such as names, places or explicit calls for help. Multimodal fusion models, such as CrisisKAN, attempt to address these gaps by combining image and text features, but rely on external knowledge sources such as Wikipedia. These sources are often unavailable during the early stages of a crisis and introduce significant computational overhead, making them unsuitable for real-time deployment.

This presents a fundamental challenge: How can we design an accurate, efficient, and scalable multimodal classification model that performs well in real-time, without relying on delayed or computationally expensive external knowledge sources? In this work, we explore whether VLMs, particularly those capable of generating semantic image captions, can serve as a lightweight and real-time alternative to Wikipedia-based knowledge infusion in multimodal learning. We formulate the problem as a multimodal classification task over image-text pairs, with the goal of predicting whether a social media post is informative (Task 1), identifying its humanitarian category (Task 2), and assessing the severity of damage (Task 3).

Additionally, we evaluated the feasibility of using a fine-tuned CLIP model as a lightweight unified encoder for both modalities. This allows us to compare traditional multimodal fusion approaches with more compact, pre-trained models that avoid explicit cross-modal fusion. Our objective is to assess the trade-offs between model complexity, accuracy, and responsiveness, in order to inform the development of real-time disaster response systems deployable under realistic computational constraints.

IV. Dataset Structure

This study uses the CrisisMMD dataset, a publicly available human-annotated collection of pairs of tweet image captured during real-world natural disasters. The dataset contains content from events such as hurricanes, earthquakes, floods, and wildfires, collected via event-specific keywords and hashtags. Each sample is paired with labels that support multiple classification tasks relevant to disaster response.

A. Composition of Dataset

CrisisMMD includes more than 17,000 pairs of text from images collected using event-specific keywords and hashtags, ensuring relevance to disaster response applications. It was sourced from seven major crisis events including Hurricane Irma, Hurricane Harvey, Hurricane Maria, the Mexico earthquake, California wildfires, the Iraq-Iran earthquake, and the Sri Lanka floods. Table 1 provides an overview of the crisis events, the keywords used for data collection, and the time period during which the data were gathered. All samples were annotated by human experts to ensure quality and consistency, with multiple validation rounds to reduce labeling errors and ambiguities.

TABLE I
CRISISMMD DATASET DETAILS INCLUDING EVENT NAMES, KEYWORDS USED FOR DATA COLLECTION, AND DATA COLLECTION PERIOD.

Crisis event	**Keywords**	**Start date**	**End date**
Hurricane Irma	*Hurricane Irma, HurricaneIrma, Irma storm,...*	Sep 6, 2017	Sep 21, 2017
Hurricane Harvey	*Hurricane Harvey, Harvey, HurricaneHarvey,...*	Aug 26, 2017	Sep 20, 2017
Hurricane Maria	*Hurricane Maria, Maria Storm, Maria Cyclone,...*	Sep 20, 2017	Nov 13, 2017
Mexico earthquake	*Mexico earthquake, mexicoearthquake,...*	Sep 20, 2017	Oct 6, 2017
California wildfires	*California fire, California wildfires,...*	Oct 10, 2017	Oct 27, 2017
Iraq-Iran earthquake	*Iran earthquake, Iraq earthquake, halabja earthquake,...*	Nov 13, 2017	Nov 19, 2017
Sri Lanka floods	*SriLanka floods, FloodSL, SriLanka flooding,...*	May 31, 2017	Jul 3, 2017

B. Classification Tasks

Each tweet image pair in the CrisisMMD dataset is labeled for three classification tasks that reflect the core information needs in the response to disasters. Task 1 focuses on identifying informative posts for prioritization. Task 2 assigns humanitarian relevance categories, while Task 3 evaluates the severity of visual damage.

TABLE II
OVERVIEW OF THE THREE CLASSIFICATION TASKS IN THE CRISISMMD DATASET, INCLUDING NUMBER OF CLASSES AND LABEL EXAMPLES.

Task	**Description**	**Classes**	**Class Labels**
Task 1	Classifies whether a tweet-image pair is informative or not.	2	Informative, Non-informative
Task 2	Identifies the humanitarian category of the post (e.g., rescue, damage).	8	Affected individuals, Injured/Dead people, Vehicle damage, Infrastructure damage, Rescue/Volunteering, Not humanitarian, etc.
Task 3	Assesses the severity of visual damage in the image.	3	Little, Mild, Severe damage

C. Dataset Splits

Due to differences in model architecture and caption availability, we used two different dataset splits in our experiments.

The first split, used for the CrisisKAN model without knowledge infusion and for the CLIP-based classifier, follows the standard provided with the CrisisMMD dataset. This version includes the full dataset and is suitable for evaluating models that do not rely on additional generated captions.

TABLE III
DATASET SPLIT USED FOR CLIP AND CRISISKAN (NO KNOWLEDGE INFUSION), BASED ON CRISISMMD SETTING A.

Task	**Train Samples**	**Validation Samples**	**Test Samples**	**Total**
Task 1	9601	4675	2724	17,000
Task 2	2874	477	451	3802
Task 3	2461	529	530	3520

For caption-infused models, a separate segment was created based on the subset of data for which image captions were successfully generated. Captioning was performed using the `phi-4` VLM by Microsoft. However, this model could only be deployed using A100 GPUs via Google Colab, which introduced both cost constraints and execution time limits. As a result, we were only able to generate captions for approximately 11,000 out of the full 17,000 image-text pairs in CrisisMMD.

The table below shows the adjusted split used for the VLM-infused models:

TABLE IV
CUSTOM DATASET SPLIT USED FOR VLM-INFUSED MODELS (CAPTION-ENHANCED), LIMITED TO SAMPLES WITH GENERATED CAPTIONS.

Task	**Train Samples**	**Validation Samples**	**Test Samples**	**Total**
Task 1	8533	1433	1434	11,400
Task 2	1500	326	347	2173
Task 3	1705	383	372	2460

This difference in data availability had a measurable impact on model performance, especially for Tasks 2 and 3, which were more sensitive to data scarcity and class imbalance. Despite this, the experiments still allowed us to effectively evaluate the feasibility of infusion of caption-based knowledge in real-time disaster classification.

D. Hypotheses

We designed our experiments to evaluate the feasibility of real-time alternatives to external knowledge infusion in multimodal crisis classification. Specifically, we formulate three core hypotheses aligned with our model configurations and the challenges of rapid deployment in resource-constrained environments.

First, we hypothesized that removing Wikipedia-based knowledge infusion from CrisisKAN would lead to performance degradation, particularly in fine-grained classification tasks such as humanitarian relevance (Task 2) and damage severity (Task 3). These tasks require deeper contextual understanding, which Wikipedia previously provided.

Second, we expected that incorporating captions generated by a VLM could serve as an effective substitute for Wikipedia-based knowledge. By appending automatically generated image captions to tweet text, we aim to recover some of the lost contextual richness, particularly by improving classification performance on the informativeness task (Task 1), where semantic cues from the image are often sufficient.

Finally, we hypothesized that fine-tuning the CLIP model as a unified encoder for both modalities would offer a lightweight and computationally efficient alternative. However, we anticipated that this model would exhibit a trade-off in classification performance due to the absence of explicit multimodal fusion mechanisms.

These hypotheses collectively reflect a broader research question at the core of our study: How can we balance classification performance, processing speed, and architectural simplicity in AI systems designed for real-time disaster response?

V. Methodology: Our Approach

This section describes the three experimental architectures implemented for crisis classification: (1) a baseline CrisisKAN model without knowledge infusion, (2) a caption-infused multimodal model using a Visual Language Model (VLM), and (3) a CLIP-based classifier. All models process tweet-image pairs and are trained for three classification tasks: informativeness, humanitarian category, and damage severity.

A. CrisisKAN without Knowledge Infusion

CrisisKAN was originally designed to incorporate external knowledge (Wikipedia) into multimodal learning. In our modified version, we removed this knowledge infusion step to evaluate the standalone fusion of visual and textual modalities. Formally, each data instance is represented as a tuple $x_i = (I_i, T_i)$, where I_i is the input image and T_i is the associated tweet. The objective is to learn a function $f : I \times T \to Y$, where Y is the set of class labels for each task. Given a dataset $(I_i, T_i, y_i)_{i=1}^{N}$, we seek to minimize prediction error for unseen tweet-image pairs.

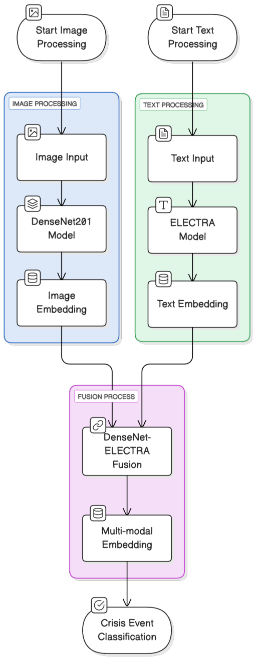


Fig. 1. Architecture of CrisisKAN without knowledge infusion.

1) Visual Feature Extraction: We use a DenseNet convolutional neural network to extract image features. For each input image I, the image encoder produces a high-level feature map:

$$\mathcal{Z}_I = \text{DenseNet}(I) \in \mathbb{R}^{c \times h \times w}$$

where c, h, and w are the number of channels, height, and width of the feature map. These features preserve spatial locality, allowing the attention module to focus on specific regions, such as flooded roads or collapsed buildings.

2) Textual Feature Extraction: The text input T is constructed by concatenating the tweet and its associated event name:

$$S = \text{Concat}([T_{\text{tweet}}, T_{\text{event}}])$$

This sequence is tokenized and passed through the ELECTRA transformer encoder:

$$[h_{S_0}, h_{S_1}, \ldots, h_{S_N}] = \text{ELECTRA}(S)$$

We use the [CLS] token embedding from the last hidden state as the textual representation:

$$\mathcal{Z}_S = h_{S_0} \in \mathbb{R}^d$$

3) MultiModal Fusion (Guided Cross Attention): We apply a guided attention mechanism that first uses self-attention to clean modality-specific noise and then cross-attention to align text and image features. The attention function is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

Self-attention is applied separately:

$$\mathcal{Z}'_I = \text{Attention}(\mathcal{Z}_I), \quad \mathcal{Z}'_S = \text{Attention}(\mathcal{Z}_S)$$

The output features are projected to a shared dimension $K = 100$:

$$\tilde{\mathcal{Z}}_I = F(W_I^T \mathcal{Z}'_I + b_I), \quad \tilde{\mathcal{Z}}_S = F(W_S^T \mathcal{Z}' S + b_S)$$

Cross-attention masks are learned as:

$$\alpha_{\mathcal{Z}_I} = \sigma(W'_I \mathcal{Z}'_S + b'_I), \quad \alpha_{\mathcal{Z}_S} = \sigma(W'_S \mathcal{Z}'_I + b'_S)$$

These are then applied to the modality features:

$$\hat{\mathcal{Z}}_I = \alpha_{\mathcal{Z}_I} \odot \tilde{\mathcal{Z}}_I, \quad \hat{\mathcal{Z}}_S = \alpha_{\mathcal{Z}_S} \odot \tilde{\mathcal{Z}}_S$$

4) Classification: The fused representations are passed to a fully connected network for classification, trained with standard cross-entropy loss:

$$\mathcal{L} = -\sum_{c \in \mathcal{Y}} y_c \log \hat{y}_c$$

B. *Caption-Enhanced VLM Infused Model*

To evaluate real-time alternatives to Wikipedia, we used the phi-4 Visual Language Model to generate semantic captions from images. Each image-caption pair was appended to the original tweet text and encoded using ELECTRA. The rest of the architecture, denseNet for visual encoding and guided cross-attention for fusion, remained the same.

This approach was trained in two configurations:

- **Single-task learning**: each model handles one task independently.
- **Multi-task learning**: a shared encoder feeds three task-specific classification heads. The total loss is computed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$$

where $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ are the cross-entropy losses for Tasks 1, 2, and 3.

Due to computational limits with Microsoft phi-4, which required access to A100 GPUs via Colab, we could only caption approximately 11,000 samples. This reduced the training data available for VLM-infused models.

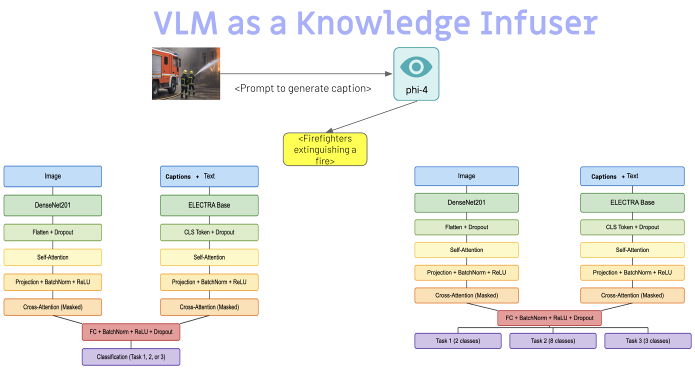
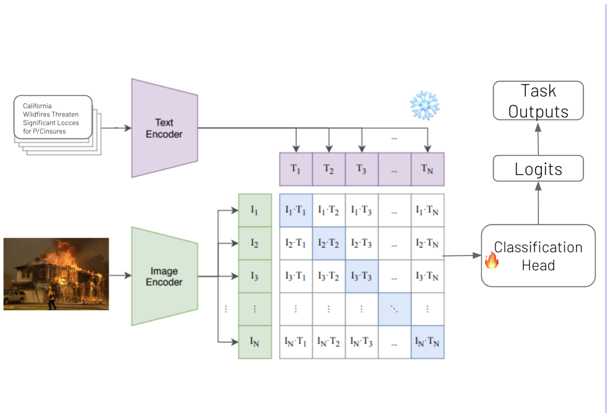


Fig. 2. Architecture of the VLM-based knowledge infusion models (single-task and multi-task).

C. *Visual Language Model as a Classifier*

We selected CLIP (Contrastive Language–Image Pretraining) as the encoder for our classifier because of its unique ability to jointly understand both textual and visual inputs in a unified embedding space. This is especially valuable in disaster response scenarios where social media posts often include a combination of text and images. Unlike traditional models that treat modalities separately and require explicit fusion layers, CLIP inherently aligns image and text representations, making it a lightweight and efficient alternative for multimodal understanding. It was pre-trained on more than 400 Million (image,text) pairs to facilitate zero shot transfer to downstream tasks. The text encoder and the image encoder project inputs to a latent space with the same number of dimensions. Given N (image and corresponding test pairs), CLIP is trained to predict which of the N * N possible (image,text) pairings across the batch occurred. CLIP learns by jointly training an image encoder and a text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs and minimize the similarity of the incorrect pairs. Training optimizes a symmetric cross entropy Loss over the similarity scores [10]. In our case we would interpret it as: given a batch of image and tweet pairs for the disaster, we calculate the cosine similarity of the actual image tweet pairs that correspond to one another and minimize the cosine similarity of the image tweet pairs that do not correspond to one another. As an addition to CLIP and to enable classification, we added a Fully connected Linear Layer to act as the classifier head and give the output based on the tasks. This approach allowed us to explore a computationally cheaper and faster alternative to models like CrisisKAN, which rely on heavyweight fusion modules and external knowledge infusion. In our use case, CLIP served as both the text and image encoder, effectively reducing the number of components in the pipeline while still capturing visual and textual semantic information necessary for classifying informativeness, humanitarian type, and damage severity. Figure 3 fully shows the training pipeline for VLM as a classifier.



Source: Alec Redford et. al Learning Transferable Visual Models From Natural Language Supervision, Feb 2021

Fig. 3. Modification of CLIP to classify. We attached a fully connected Linear Layer to enable classification for the three tasks in CrisisMMD dataset.

It is important to note that during training, all parameter layers in the CLIP model were frozen and only the classification head's parameters were trainable.

D. *Training Configuration*

We tested five experimental configurations across our three main modeling strategies: the CrisisKAN baseline, VLM-infused models (both single-task and multi-task), and VLM-based classifiers (fine-tuned CLIP). Each experiment varied in batch size, learning rate scheduling, and compute resources. The table below summarizes the experimental setup for each configuration.

TABLE V
EXPERIMENTAL CONFIGURATIONS FOR VLM-BASED CRISIS CLASSIFICATION

Setting	Epochs	Batch	Optimizer	Scheduler	Compute	Time/Epoch
Baseline (CrisisKAN)	50	64	Adam	–	NVIDIA A100 40GB	–
VLM Infuser (single task)	20	32	AdamW	ReduceLR / SGD	Colab A100 40GB	45 min
VLM Infuser (multi-task)	20	32	AdamW	ReduceLR Plateau	Colab A100 40GB	2 hrs
VLM Classifier (1 layer)	6	16	AdamW	Constant LR	PSC V100 32GB	1 min
VLM Classifier (2 layers)	20	16	AdamW	Constant LR	V100	3 mins

VI. RESULTS AND DISCUSSION

The table below summarizes the performance of all model variants across the three tasks in CrisisMMD. Our results indicate that while CrisisKAN (with Wikipedia-based knowledge infusion) sets a strong performance benchmark, alternatives based on Visual Language Models (VLMs) show promising results, especially in terms of flexibility and inference feasibility in real-time settings.

TABLE VI
MODEL PERFORMANCE ACROSS THREE CRISIS CLASSIFICATION TASKS

Model \Task	Task 1	Task 2	Task 3
CrisisKAN	Acc: 86.9 F1: **86.2%**	Acc: 90.1% F1: 90.9%	Not reported
CrisisKAN (w/o knowledge infusion)	Overall Acc: **70.0%**	–	–
VLM Knowledge Infuser (Single task)	Acc: 75.59% F1: 75.21%	Acc: 56.78% F1: 56.68%	Acc: 63.98% F1: 60.48%
VLM Knowledge Infuser (Multi-task)	Acc: 79.43% F1: 79.40%	Acc: 63.41% F1: 61.78%	Acc: 67.47% F1: 59.46%
VLM Classifier (1-layer)	Acc: 73.89% F1: 70.70%	Acc: 50.78% F1: 41.12%	Acc: 62.76% F1: 48.4%
VLM Classifier (2-layer, fine-tuned)	Acc: 70.76% F1: 68.34%	Acc: 53.24% F1: 51.64%	Acc: 58.60% F1: 53.28%

1) Task specific Analysis: **Task 1: Informative and Non Informative** CrisisKAN (with knowledge fusion) gives the best performance with a weighted F1 score of 86.2 likely due to the rich feature infusion from Wikipedia which provided better textual context for Binary classification. On the same task, using a Visual Language model (multi-task) for knowledge infusion (mainly from captions describing the image) performs slightly higher than the VLM (single task) by $\approx 4\%$. This is a slightly lower but descent performance when compared to CrisisKAN suggesting that the knowledge that was infused using VLM may not be as dense as the knowledge infused from Wikipedia. VLM as a classifier lags behind the knowledge infusers especially when finetuned, but still offer reasonably efficient alternative. This means that CLIP's strong pretraining in addition to a simple linear layer can effectively separate binary classes. Interestingly fine-tuning slightly degraded performance possibly due to overfitting or poor hyperparameters.

Task 2: Humanitarian Category is a multiclass classification task distinguishing types of humanitarian needs (eg, vehicular, infrastructure damage, rescue or aid). CrisisKAN achieves superior performance(both in accuracy and F1 score) as well on this task showing strong multi modal comprehension and ability to distinguish between multiple classification tasks. VLM as a knowledge infuser gets a poorer performance as compared to CrisisKAN (with and without knowledge infusion). Although we get a better score with multitask, the overall score is Lower than the benchmark by far suggesting poor generalization in multiclass classification tasks. VLM as a classifier struggles in this task as well. We could attribute it the inability of the frozen encoder to learn to generalize across multiple classes and that the classifier is too shallow to learn nuanced differences in the humanitarian categories.

Task 3: Severity of Damage focused on assessing the visible or textual degree of damage, Multitask VLM achieves the best F1: 59.46 demonstrating the unique ability of joint training to recognize visua/textual cues for severity. The fine tuned VLM classifier improves over the baseline showing the benefit of finetuning to adapt the model for nuanced tasks. Task 3 benefited the most from finetuning as compared to task 1 and 2. This means that this task could benefit from end to end training of the frozen encoder.

Overall, although our experiments fail to give a performance as high as CrisisKAN, they confirm that caption-based knowledge infusion using VLMs is a viable, alternative to Wikipedia infusion. It demonstrates effectiveness particularly in simpler tasks (e.g., informativeness) and remains promising for scalable real-time use. Our experiments also show the potential of using a VLM as a joint encoder in the need for a light weight and fast way of encoding textual and visual features.

VII. CONCLUSION

This study explored practical and light weight alternatives to knowledge-infused multimodal crisis classification, using Pretrained Visual Language Models. Our experiments confirmed that while CrisisKAN with Wikipedia based knowledge infusion attains the best performance, it poses a challenge when one needs real time information about the disatser on the ground. There for, using VLM-based models with caption-based text enrichment in both single-task and multi-task settings present a reliable way of disaster classification with a good performance. Using unified CLIP encoders, also presents viable light weight and fast solutions for inference in disaster settings. VLM as a knowledge infuser demonstrated its power to capture semantic richness in images and provide a good performance for binary tasks but seem to struggle in multiclass tasks due to the limitations of class imbalance. The CLIP-based classifier, although less accurate overall, showed promise as a fast and modular alternative for resource constrained crisis enviromnents especially when lightly finetuned. Our findings serve as a foundation for research in real time, knowledge efficient and light weight crisis classification for disaster response .

VIII. Future Work

The results highlight several directions for future exploration. First, performance gaps across all tasks, especially in Task 2, suggest the need for improved class balancing techniques, potentially via weighted loss functions or augmentation. The overfitting observed in fine-tuned models could be addressed through advanced regularization, cross-validation, and architecture pruning.

We also plan to explore alternative VLMs that support image-conditioned generation to yield more context-rich captions, further narrowing the gap with Wikipedia-based infusion. For CLIP-based methods, tuning the encoder layers (rather than freezing them entirely) may help capture more nuanced cross-modal relationships.

Finally, we will integrate Named Entity Recognition (NER) to extract location references from tweets and captions, enabling geospatial situational awareness—an essential requirement in real-time crisis response scenarios.

1) Contribution: Christine helped develop the experiments for the baseline and the VLM as a classifier pipeline. Keriane developed the VLM as a knowledge infuser pipeline by gathering all captions from images and concatenating them to the textual inputs and performed training of the same model. Both Christine and Keriane equally contributed to the documentation of all the sections of the final report.

Project Resources

The following link indicate all the experiments and the configurations to replicate our project:

- **GitHub Repository:** https://github.com/Kenza1525/AI-Powered-Disaster-Response.git
- **Weights & Biases Dashboard:** https://wandb.ai/principle-paper?shareProfileType=copy

References

[1] F. Alam, F. Ofli, and M. Imran, “CrisisMMD: Multimodal Twitter Datasets from Natural Disasters,” May 02, 2018, arXiv: arXiv:1805.00713. doi: 10.48550/arXiv.1805.00713.

[2] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, ”AIDR: Artificial Intelligence for Disaster Response,” in Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 159-162.

[3] S. Gupta, N. Saini, S. Kundu, and D. Das, “CrisisKAN: Knowledge-infused and Explainable Multimodal Attention Network for Crisis Event Classification,” arXiv.org, 2024. https://arxiv.org/abs/2401.06194.

[4] S. K. Balasubramanian, V. V. Kumar, A. Sahoo, and S. Gholston, ”Mitigating Healthcare Supply Chain Challenges Under Disaster Conditions: A Holistic AI-based Analysis of Social Media Data,” International Journal of Production Research, forthcoming.

[5] E. Irwansyah, H. Young, and A. A. S. Gunawan, “Multi Disaster Building Damage Assessment with Deep Learning using Satellite Imagery Data,” International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 1, Art. no. 1, Jan. 2023.

[6] M. Abavisani, L. Wu, S. Hu, J. Tetreault, and A. Jaimes, “Multimodal Categorization of Crisis Events in Social Media,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2020, pp. 14667–14677. doi: 10.1109/CVPR42600.2020.01469.

[7] J. B. Houston, J. Hawthorne, M. F. Perreault, E. H. Park, M. Goldstein Hode, M. R. Halliwell, S. E. Turner McGowen, R. Davis, S. Vaid, J. A. McElderry, et al., ”Social media and disasters: A functional framework for social media use in disaster planning, response, and research,” Disasters, vol. 39, no. 1, pp. 1–22, 2015.

[8] Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020, March 23). Electra: Pre-training text encoders as discriminators rather than generators. arXiv.org. https://arxiv.org/abs/2003.10555v1

[9] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, “Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters,” Remote Sensing of Environment, vol. 265, p. 112636, Nov. 2021, doi: https://doi.org/10.1016/j.rse.2021.112636.

[10] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” arXiv:2103.00020 [cs], Feb. 2021, Available: https://arxiv.org/abs/2103.00020

[11] H. Ritchie, P. Rosado, and M. Roser, “Natural Disasters,” Our World in Data, Dec. 07, 2022. Available: https://ourworldindata.org/natural-disasters