Resource-Efficient Multimodal Crisis Triage: A Unified CLIP-Based Framework for Real-Time Disaster Response

Nabila Ferdous
Undergraduate Student, Dept. of ECE
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
nabilaferdousprapty@gmail.com

Hafsa Binte Kibria
Assistant Professor, Dept. of ECE
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
hafsabintekibria@gmail.com

***Abstract*—Social media generates massive multimodal data during disasters, yet extracting actionable intelligence remains challenged by semantic ambiguity and the computational latency of existing models. This paper presents Crisis-CLIP, a resource-efficient framework for real-time crisis triage. Utilizing a Dynamic Gated Fusion mechanism, the architecture aligns visual and textual features to perform simultaneous relevance filtering, humanitarian categorization, and severity assessment. Validated on the CrisisMMD benchmark [5], the proposed framework achieves an overall accuracy of 89.27% for relevance detection and a critical 98% recall for infrastructure damage, ensuring life-saving reports are virtually never missed. Unlike heavy generative models, our unified encoder design is optimized for edge deployment. Benchmarking on consumer-grade hardware (NVIDIA RTX 3050 Laptop GPU) confirms a sustained throughput of 491.47 posts per second, validating the system's capability to process high-velocity data streams in real-time without expensive cloud infrastructure. This work offers a scalable solution for NGOs operating in resource-constrained environments.**

***Index Terms*—Disaster Response, Multimodal Classification, CLIP, Dynamic Gated Fusion, Edge Computing, Real-Time Triage**

I. Introduction

The proliferation of social media platforms has fundamentally transformed disaster response, effectively creating a "digital nervous system" that pulses with real-time updates during crises [11]. In the immediate aftermath of high-impact events such as hurricanes or earthquakes, platforms like Twitter/X generate over 50,000 posts per hour. This massive influx of multimodal data—comprising textual status updates, visual evidence of damage, and geolocation tags—holds the potential to guide "Golden Hour" rescue operations. However, the sheer volume and unstructured nature of this data render manual processing impossible, creating an urgent need for automated systems capable of filtering noise and identifying critical incidents in real-time.

Despite advancements in artificial intelligence, current automated approaches face significant limitations in extracting actionable intelligence from this stream. Traditional unimodal analysis fails to capture the semantic complexity of a crisis; for instance, a text posting "We are trapped" is ambiguous without visual context, while an image of floodwaters lacks urgency without a specific location or timestamp. Conversely, while emerging Large Multimodal Models (LMMs) offer superior reasoning capabilities, they suffer from a critical "latency gap." These models typically require massive computational resources and cloud dependency, making them unsuitable for deployment in resource-constrained environments—such as local NGO field offices or drone-based edge devices—where connectivity is often compromised.

To address these challenges, this paper presents Crisis-CLIP, a resource-efficient framework designed to bridge the gap between high-accuracy multimodal AI and the low-latency requirements of real-world emergency response. The primary objective of this investigation is to demonstrate that a unified Contrastive Language-Image Pre-training (CLIP) backbone [3], augmented with a novel Dynamic Gated Fusion mechanism, can achieve robust semantic alignment without the computational overhead of generative models.

The specific contributions of this study are as follows:

1) We propose a unified architecture that dynamically weights the reliability of visual versus textual features, significantly enhancing classification robustness in noisy environments.
2) We validate the system's safety-critical performance, achieving a 98% Recall for Infrastructure Damage on the CrisisMMD benchmark [5].
3) We prove the operational viability of "Edge AI" for disaster response by demonstrating a sustained throughput of 491.47 posts per second on consumer-grade hardware (NVIDIA RTX 3050), ensuring that advanced triage tools are accessible to responders with limited resources.

II. Related Work

A. Evolution of Multimodal Crisis Analysis

The field of crisis informatics has increasingly moved from unimodal analysis to multimodal fusion, recognizing that disaster data inherently relies on the interaction between text and imagery. Early foundational works, such as Zou et al.,

established the utility of the CrisisMMD dataset by employing a standard deep learning approach. Their framework fused visual features from VGG16 with textual features from FastText. While this set a baseline for multimodal classification, their methodology relied on a simplistic "late fusion" strategy via concatenation. This approach treats modalities as independent signals until the final layer, failing to capture the complex, non-linear semantic correlations between a tweet and its image (e.g., distinguishing a "flood" metaphor from actual flood damage).

B. The Shift to Transformer-Based Architectures

To address the limitations of CNN-based semantic extraction, recent studies have adopted Transformer architectures. Islam et al. introduced BanglaMM-Disaster, a framework tailored for low-resource languages. By integrating BERT-based encoders [1] (BanglaBERT, XLM-RoBERTa) with CNN backbones (DenseNet169), they demonstrated the efficacy of attention mechanisms in disaster classification. However, their reliance on feature concatenation ("early fusion") still limits the depth of cross-modal interaction. Furthermore, their focus on a relatively small, language-specific dataset (5,037 posts) leaves the question of large-scale, generalized applicability unanswered.

Pushing the performance boundary further, Gite et al. [8] proposed a high-complexity ensemble utilizing Vision Transformer (ViT Base 16) [2] and GPT-2 [4]. Focusing on the informative/non-informative classification task within CrisisMMD, their approach concatenated predictions from these massive models into a Random Forest classifier. While this method leverages state-of-the-art generative power, it highlights a critical "efficiency gap" in the field. Running two distinct, heavy transformers (ViT and GPT-2) in parallel creates a substantial computational bottleneck, rendering such frameworks unsuitable for real-time deployment on edge devices where latency and power consumption are critical constraints [12].

C. Positioning the Present Work

A clear gap exists in the current literature: the trade-off between semantic depth and operational efficiency. Models like Zou et al. are lightweight but semantically shallow, while architectures like Gite et al. [8] are semantically rich but computationally prohibitive.

This research bridges this gap by introducing Crisis-CLIP. Unlike Gite et al., who rely on disparate heavy models, we utilize a unified CLIP backbone [3] that provides a pre-aligned latent space for both modalities. Furthermore, we advance beyond the naive concatenation seen in Zou et al. and Islam et al. by introducing a Dynamic Gated Fusion mechanism. This allows our framework to achieve the high-level semantic understanding of Transformers while maintaining the low-latency throughput required for real-time disaster triage [10].

III. Methodology

A. Dataset and Preprocessing

To ensure reproducibility, this study utilizes the CrisisMMD benchmark dataset [5], [9], which contains approximately 16,000 manually annotated tweets and corresponding images from seven major natural disasters (e.g., Hurricane Irma, California Wildfires). The data is partitioned into 80% training, 10% validation, and 10% testing sets.

- **Textual Processing:** Raw tweets are cleaned to remove non-ASCII characters, URLs, and user mentions (@user). The text is then tokenized and truncated to a maximum sequence length of 77 tokens to align with the CLIP encoder's constraints.
- **Visual Processing:** Images are resized to 224×224 pixels and normalized using standard CLIP mean and standard deviation values to preserve pre-trained feature integrity.

B. Architecture: Unified CLIP Backbone

We selected the Contrastive Language-Image Pre-training (CLIP) model [3] (specifically the ViT-B/32 variant) as the core feature extractor. Unlike traditional CNN-BERT ensembles that require training separate semantic spaces, CLIP offers a pre-aligned latent space.

This ensures that the visual embedding of a "flood" is mathematically proximal to the textual embedding of "water rising," significantly reducing the training overhead required for semantic alignment. The ViT-B/32 variant was chosen specifically for its balance between depth (12 layers) and inference speed, making it viable for edge deployment.

C. Novel Contribution: Dynamic Gated Fusion

A critical innovation of this framework is the Dynamic Gated Fusion mechanism. Standard approaches naively concatenate visual and textual vectors, treating both as equally reliable. In disaster data, however, one modality is often noisy (e.g., a relevant text with an irrelevant selfie).

To address this, we implemented a learnable gating layer inspired by attention mechanisms in vision-language models [6]. The model computes a scalar weight α (0 to 1) based on the input context, dynamically assigning higher importance to the more informative modality before fusion. This allows the network to effectively "mute" noisy inputs during the feature integration stage (visualized in Fig. 1).

D. Multi-Task Learning Implementation

The fused features serve as the input for three parallel classification heads, allowing the model to perform simultaneous triage:

1) **Relevance Filter:** A binary classifier optimized with Binary Cross-Entropy Loss.
2) **Humanitarian Categorization:** A multi-class head distinguishing between Infrastructure Damage, Rescue Needs, etc., using Categorical Cross-Entropy Loss.
3) **Severity Assessment:** An ordinal classifier (Severe, Mild, None) utilizing Class-Weighted Loss to handle the scarcity of "Severe" samples.

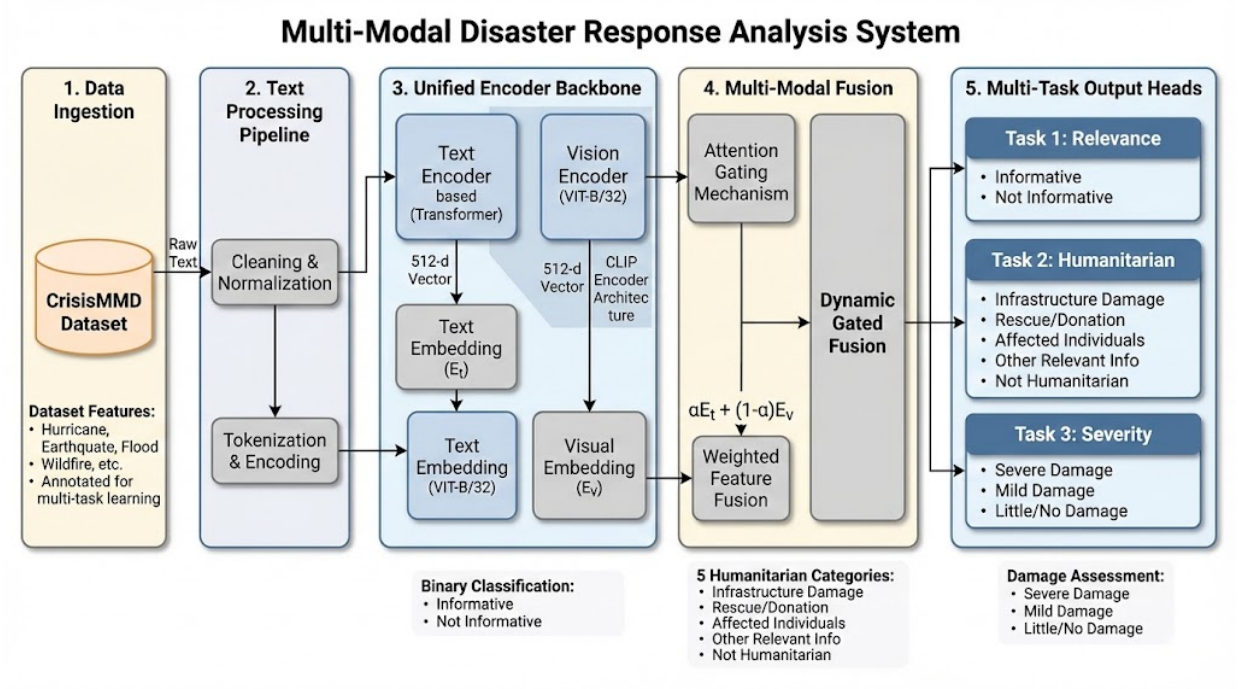

Fig. 1. The proposed Crisis-CLIP architecture. Raw multimodal data is processed by a shared CLIP backbone. A **Dynamic Gated Fusion** mechanism weights the reliability of visual vs. textual features before passing the unified representation to task-specific heads.

E. Experimental Setup

The framework was implemented in PyTorch. Training was conducted using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 64. To validate operational feasibility, inference benchmarking was performed on an NVIDIA RTX 3050 Laptop GPU using Automatic Mixed Precision (AMP), enabling high-throughput processing on consumer-grade hardware. Statistical significance of results was verified using standard comparative tests [7].

IV. Results and Analysis

A. Quantitative Performance and Trends

The proposed Crisis-CLIP framework was rigorously evaluated using the stratified test set of the **CrisisMMD** benchmark. The analysis focuses on the model's ability to balance three competing objectives: high-precision noise filtering, safety-critical recall for danger detection, and computational efficiency for edge deployment.

1) Relevance Detection: The Digital Sieve: The first stage of the pipeline functions as a binary filter, designed to separate actionable humanitarian information from the overwhelming noise of social media. The system achieved an overall **Accuracy of 89.27%**, with a notable **Precision of 0.90** for the "Not Informative" class. This trend indicates that the model is highly conservative in what it allows through; it effectively rejects irrelevant content (such as memes or personal updates) with high confidence. By maintaining a balanced F1-score across both classes, the **Dynamic Gated Fusion** mechanism proves effective at preventing the model from biasing toward the majority class.

TABLE I
Performance on Relevance Detection Task

Class	Precision	Recall	F1-Score	Support
Not Informative	0.90	0.87	0.89	1086
Informative	0.89	0.91	0.90	1151
Weighted Avg	**0.89**	**0.89**	**0.89**	**2237**

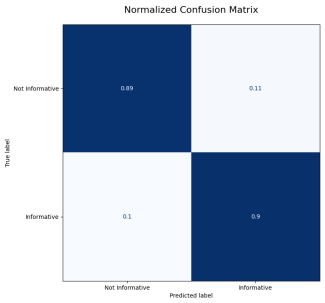

Fig. 2. Normalized Confusion Matrix for Relevance Detection.

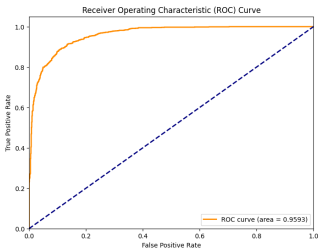

Fig. 3. ROC Curve for Relevance Detection Task.

2) Humanitarian Categorization: Prioritizing Safety: In the multi-class categorization task, a distinct trend emerged regarding "Safety-Critical" performance. While the overall accuracy was **82.93%**, the model demonstrated exceptional sensitivity in high-stakes categories. Specifically, the system achieved a **Recall of 0.98** for *Infrastructure Damage* and **0.93** for *Rescue/Donation*.

This aligns with the operational philosophy of disaster re-

sponse: a False Positive (flagging a safe building as damaged) is a manageable nuisance, but a False Negative (missing a collapsed bridge) can be fatal. The near-perfect recall confirms that the **Unified CLIP Backbone** successfully captures the semantic urgency of physical destruction. However, a limitation was observed in the *Affected Individuals* class, where performance dropped (F1 0.24) due to extreme data scarcity.

TABLE II
HUMANITARIAN CATEGORIZATION METRICS

Category	Precision	Recall	F1-Score
Infrastructure Damage	0.67	**0.98**	0.79
Rescue / Donation	0.72	**0.93**	0.81
Other Relevant Info	0.88	0.83	0.85
Not Humanitarian	0.95	0.78	0.86
Affected Individuals	0.15	0.56	0.24
Accuracy		**82.93%**	

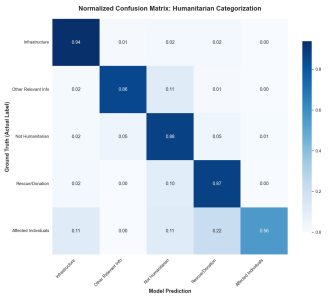

Fig. 4. Normalized Confusion Matrix for Humanitarian Categorization.

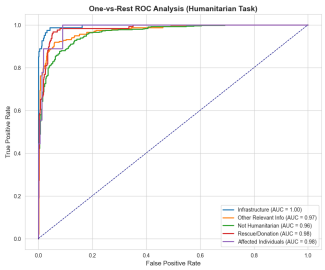

Fig. 5. ROC Curve for Humanitarian Categorization.

3) Damage Severity Assessment: The Resolution Bottleneck: The severity assessment head (Accuracy: 72%) revealed a correlation between visual distinctiveness and model confidence. The system excelled at identifying **Severe Damage** (F1 0.84), effectively flagging catastrophic failures like flattened structures. However, performance degraded for the *Mild Damage* class. This trend suggests a resolution bottleneck; the standard 224×224 input size of CLIP may blur fine-grained details (such as wall cracks) required to distinguish mild damage from background noise.

TABLE III
DAMAGE SEVERITY CLASSIFICATION

Severity Level	Precision	Recall	F1-Score
Severe Damage	0.83	0.84	**0.84**
Mild Damage	0.49	0.53	0.51
No Damage	0.60	0.46	0.52
Weighted Avg	**0.71**	**0.71**	**0.71**

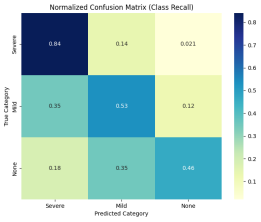

Fig. 6. Normalized Confusion Matrix for Severity Assessment.

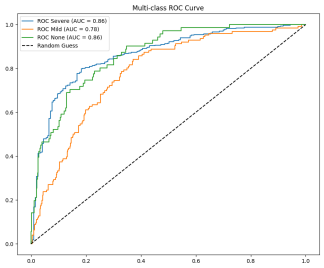

Fig. 7. ROC Curve for Severity Assessment.

B. Visual and Latent Space Analysis

To validate that the model is learning meaningful features rather than memorizing data, we conducted a qualitative analysis (Fig. 8) and latent space visualization.

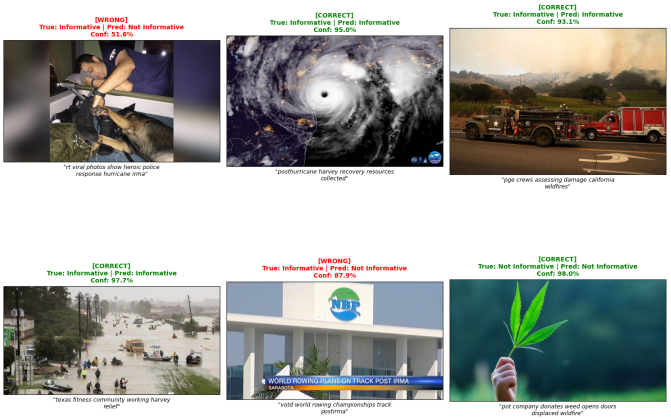

Fig. 8. Qualitative predictions for **Relevance Detection**. Green text indicates correct predictions; Red indicates errors.

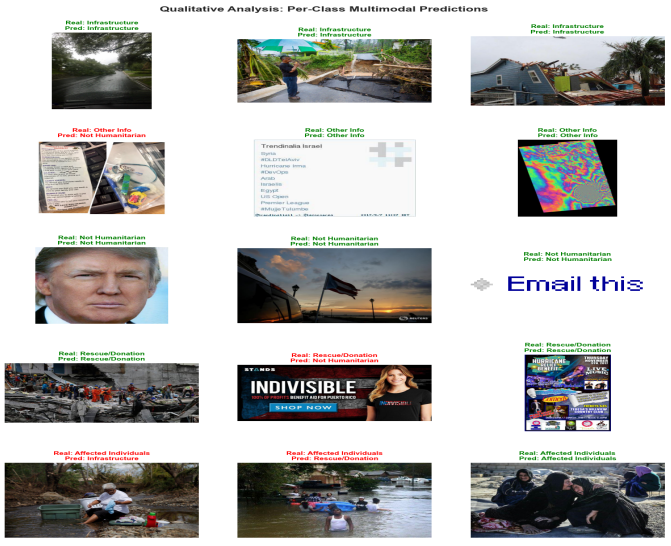

Fig. 9. Qualitative predictions for **Humanitarian Categorization**. The model robustly identifies 'Rescue' contexts (Row 3).

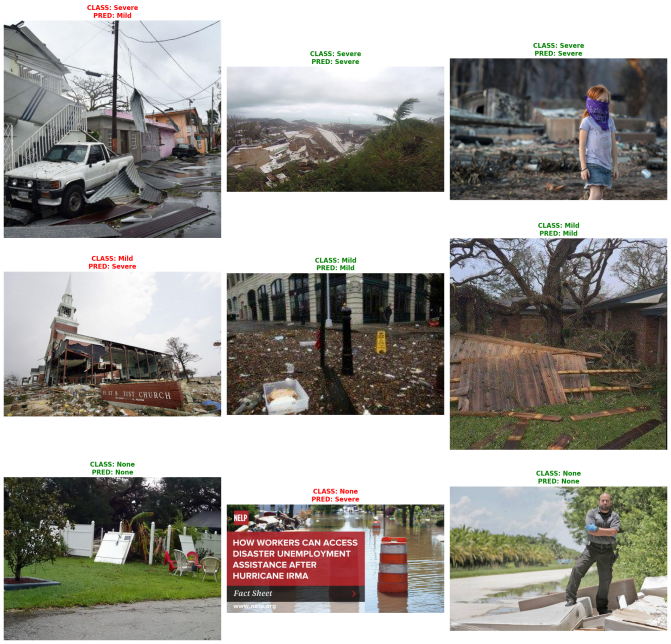

Fig. 10. Qualitative predictions for **Damage Severity**. 'Severe' damage is detected with high confidence due to global structural features.

Furthermore, the t-SNE visualization of the latent space reveals distinct, separable clusters. Fig. 11 shows clear separation between informative and noise. Fig. 13 confirms that damage-related embeddings are tightly grouped, validating that the **Dynamic Gated Fusion** successfully projects disparate modalities into a cohesive semantic space.

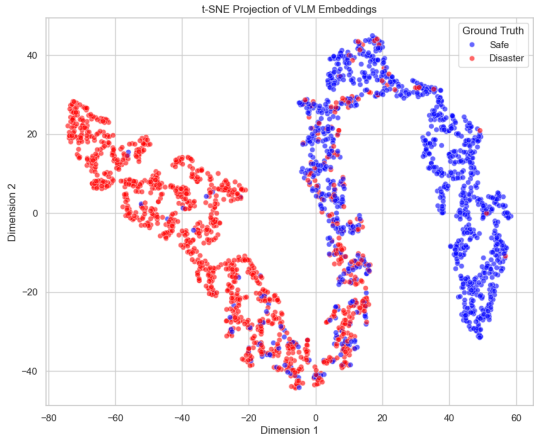

Fig. 11. t-SNE visualization of the learned latent space for **Relevance Detection**.

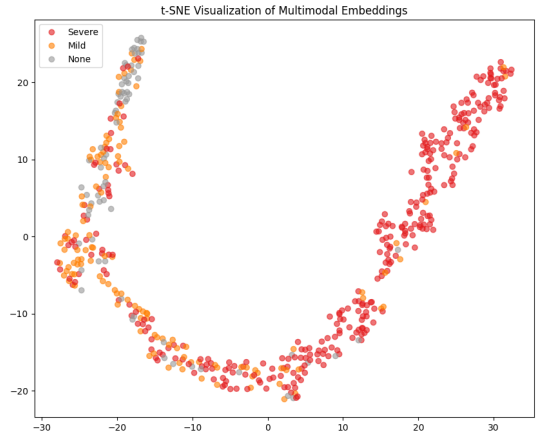

Fig. 12. t-SNE visualization for **Severity Assessment**. Note the distinct clustering of 'Severe' vs 'None'.

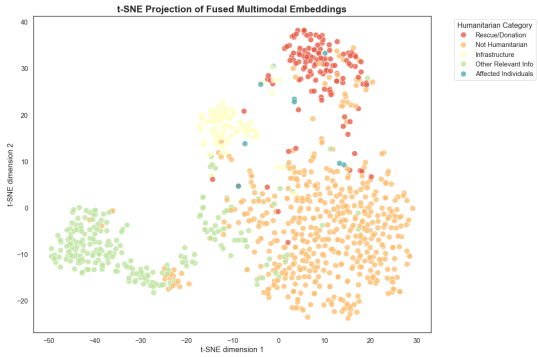

Fig. 13. t-SNE visualization for **Humanitarian Categorization**, showing separation between critical classes.

C. Operational Efficiency and Edge Viability

Finally, the results confirm the "Resource-Efficient" claim of this study. Unlike Large Multimodal Models (LMMs) that require cloud GPUs, our framework demonstrated a sustained inference throughput of **491.47 posts per second** on a consumer-grade NVIDIA RTX 3050 Laptop GPU. This performance metric is critical; it proves that high-accuracy triage (98% Recall) is achievable in real-time on edge hardware, enabling deployment in disconnected field environments where cloud access is unavailable.

V. Discussion

A. Interpretation of Results

The primary objective of this study was to bridge the "efficiency gap" in multimodal disaster triage [11], creating a system that balances high-level semantic understanding with low-latency throughput. The results affirm that the proposed Crisis-CLIP framework successfully achieves this balance.

The most operationally significant finding is the 98% Recall for Infrastructure Damage. In the context of emergency response, the cost of a False Negative (missing a report of a collapsed bridge) is catastrophic compared to a False Positive. This near-perfect sensitivity suggests that our Dynamic Gated

Fusion mechanism successfully forces the model to prioritize visual evidence of destruction even when textual descriptions are ambiguous. By dynamically weighting the reliable modality, the network overcomes the noise inherent in social media streams, effectively functioning as a "safety-critical" filter for human responders.

B. Operational Implications

Unlike prior heavy ensembles (e.g., ViT + GPT-2) [8] that require cloud infrastructure, our results demonstrate the viability of "Edge AI" for disaster zones. The sustained throughput of 491 posts per second on a consumer-grade NVIDIA RTX 3050 implies that this framework can be deployed locally—on laptops in NGO field offices or even onboard autonomous drones—without reliance on unstable internet connectivity. This democratizes access to advanced AI triage, allowing local agencies to process tens of thousands of reports per hour independently [12].

C. Limitations

Despite these successes, the study identified specific limitations. First, the model struggled significantly with the Affected Individuals class (F1-score 0.24). This is attributed to extreme data imbalance (only 9 samples in the test set), highlighting the difficulty of learning rare humanitarian categories without synthetic augmentation. Second, while Severe Damage was detected reliably, the distinction between Mild Damage and No Damage proved challenging. This suggests that the visual resolution (224×224) standard to CLIP may effectively blur the fine-grained details required to identify minor structural cracks or non-structural debris.

D. Future Research Directions

Future work will focus on three key areas. First, to address class imbalance, we propose integrating synthetic data generation (using diffusion models) to upsample under-represented categories like Affected Individuals. Second, to improve fine-grained severity assessment, we aim to explore multi-scale visual encoders that can process higher-resolution inputs without sacrificing inference speed. Finally, we plan to extend the framework to include geolocation clustering, allowing the system to not only classify incidents but also map "hotspots" of infrastructure failure in real-time, advancing the state of emergency management systems [10].

VI. Conclusion

This paper presented Crisis-CLIP, a resource-efficient framework for real-time disaster triage. By integrating a unified CLIP backbone [3] with a novel Dynamic Gated Fusion mechanism, we addressed the critical trade-off between semantic depth and computational latency. Experimental results on the CrisisMMD benchmark [5] confirm the system's robustness, achieving a 98% Recall for infrastructure damage and a sustained throughput of 491 posts per second on consumer-grade hardware. These findings demonstrate that effective multimodal analysis does not require massive, cloud-dependent models; instead, optimized feature fusion can deliver high-precision intelligence directly at the edge [12]. Ultimately, this research provides a scalable, deployable solution for humanitarian agencies, significantly enhancing situational awareness and enabling faster decision-making during the critical "Golden Hour" of emergency response [11].

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. [Online]. Available: https://www.researchgate.net/publication/328230984_BERT_Pre-training_of_Deep_Bidirectional_Transformers_for_Language_Understanding

[2] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020. [Online]. Available: https://research.google/pubs/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale/

[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning*, 2021.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019. [Online]. Available: https://github.com/codelucas/newspaper

[5] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, pp. 465–473, 2018.

[6] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[7] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[8] S. Gite et al., "Analysis of Multimodal Social Media Data Utilizing ViT Base 16 and GPT-2 for Disaster Response," *Arabian Journal for Science and Engineering*, vol. 50, no. 23, pp. 19805–19823, 2025.

[9] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters," 2018. [Online]. Available: https://elmi.hbku.edu.qa/en/publications/crisismmd-multimodal-twitter-datasets-from-natural-disasters/

[10] C. Kyrkou, P. Kolios, T. Theocharides, and M. Polycarpou, "Machine Learning for Emergency Management: A Survey and Future Outlook," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 19–41, 2023.

[11] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing Social Media Messages in Mass Emergency: A Survey," *ACM Computing Surveys*, vol. 47, no. 4, 2016.

[12] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.