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

Abstract

In spite of the massive data being collected through social media during disasters, it becomes a challenge in garnering actionable intelligence from the data since most of the extracted information falls into resides due to semantic complexity and computational difficulties.Word sense disambiguation tends to fall into two categories: methods weak in semantics (like concatenation) or those needing great computing resources (like big multimodal transformer models).. This paper proposes Crisis-CLIP, a resource-efficient approach for crisis situation management that bridges the accuracy–efficiency gap. Built upon a unified CLIP backbone, the proposed model includes a novel "Dynamic Gated Fusion" mechanism, which weighs the features of images and text in a flexible way based on their reliability. This architecture ensures a strong classification capacity even with noisy and unstructured social media environments. Evaluation on the CrisisMMD benchmark dataset results in Crisis-CLIP obtaining an accuracy of 89.27%. Notably, it supports a recall level of 98% for detecting damage to infrastructure, which ensures that life-critical are rarely missed. Most importantly, the framework processes 491.47 posts per second on consumer-grade hardware (NVIDIA RTX3050 LaptopGPU), resulting in a speedup of 25 times compared to transformer-based ensemble methods while maintaining superior accuracy. This operational efficiency enables real-time deployment in disaster zones without any reliance upon cloud infrastructure.

Keywords: Disaster Response, Multimodal Classification, CLIP, Dynamic Gated Fusion, Edge Computing, Real-Time Triage

1 Introduction

Social media sites have revolutionized the way in which disaster response procedures are handled by enabling real-time situational awareness during disaster cases. In the wake of a disaster such as a hurricane or an earthquake, for example, social media generates more than 50,000 posts per hour [5, 6], which entails a huge volume of textual data, visual content, and geographic metadata. When there are events like hurricanes or earthquakes, people post a lot on Twitter and other platforms. In fact, people post over 50,000 items per hour. This is a lot of information including what people are saying, pictures of the damage, and where they are. All of this information can really help us during the hour after a disaster, which is a very important time for rescue operations. Social media can help guide these rescue operations during this time often called the "Golden Hour" [7, 8]. Still, the amount of data and its unstructured form make processing it by hand not doable. There's a big need for automated systems to filter noise and find important incidents as they happen.[9].

While there has been significant progress in artificial intelligence, processing and obtaining actionable intelligence from social networks during disasters is computationally prohibitively expensive. Even unimodal techniques using only text or imagery reach an accuracy of merely 65% to 72% on crisis-related datasets [11], while large multimodal models like Large Multimodal Models (LMMs)—GPT-4V—result in latency times that are between 15 and 30 times longer for applications [10], which is a latency bottleneck for processing crisis-related social networks. In the wake of Hurricane Harvey, which occurred in 2017, there were over 50,000 tweets per hour, but existing ensemble techniques, like ViT+GPT-2 [10], are only able to process 15-20 posts per second on cloud infrastructure and take over 45 minutes to process, which is far beyond the "Golden Hour" timeframe of a crisis's disaster response time [6].

To overcome these challenges, Crisis-CLIP was introduced in this paper, a resource-efficient framework that aims to cover the gap between high accuracy of multimodal AIs and the requirements of real-world emergency response in terms of low latency. The primary objective of the current investigation is to illustrate that a unified Contrastive Language-Image Pre-training (CLIP) backbone [1], equipped with a new Dynamic Gated Fusion mechanism, can successfully achieve semantic alignment without the computational expense of generative models.

The specific contributions of this study are as follows:

1. A new architecture is proposed where feature integration is based on dynamic weighting depending upon the reliability of visual and textual information, thus improving the robustness of classification.

2. The safety-critical performance of the proposed system is evaluated, and a 98% recall is reported for Infrastructure Damage using the CrisisMMD dataset.[4].
3. The practical viability of Edge AI in disaster response is demonstrated by reaching a throughput of 491.47 posts per second using consumer-grade hardware (NVIDIA RTX 3050), ensuring that high-end triaging tools are accessible to those with limited resources.[14, 15].

2 Related Work

2.1 Evolution of Multimodal Crisis Analysis

Today, crisis informatics has shifted from unimodal analysis to multimodal fusion research. Past research by Zou et al. demonstrated the efficiency of fusion between visual features (using VGG16) and textual features (using FastText), by employing concatenation in a later stage of fusion. Crisis informatics understands that when something bad happens, like a disaster, we get information from what people write and from pictures [7, 8]. Some early studies, like the ones done by Zou et al. [11], showed how useful the CrisisMMD dataset can be. They used an approach with deep learning. They combined features extracted from pictures using VGG16 [**?**] with features extracted from text using FastText. This created a starting point for looking at multiple types of information together to make decisions. The way they did it was simple: they just combined the information from the pictures and the text at the end. This approach treats modalities as separate signals until the final layer [12]. It does not capture complex, non-linear connections between a tweet and its image. For instance, it cannot differentiate between a "flood" metaphor and actual flood damage [**?**].

2.2 The Shift to Transformer-Based Architectures

To overcome the limits of CNN-based methods, recent studies have shifted to Transformer architectures. [3]. For example, Islam et al. introduced "BanglaMM-Disaster," a framework designed for low-resource languages by combining BanglaBERT and XLM-RoBERTa with DenseNet169[**?**]. While they showed that attention mechanisms work well [16] their use of late fusion through simple concatenation restricts deep interaction between different types of data during feature learning. This limitation hinders cross-modal semantic alignment. Additionally, the study depended on a small, language-specific dataset of 5,037 posts, which affects the framework's ability to generalize.

Gite et al.[10] reviewed the content potential of the CrisisMMD dataset by proposing a heavy ensemble architecture which incorporates Vision Transformer (ViT-Base16) [2] and GPT-2[**?**] in an attempt to improve classification performance. They were able to effectively utilize the semantic advantages of both modalities by feeding the predictions of these different models into a Random Forest classifier. This method, however, shows a major computational obstruction. Two large transformer models must be operated in parallel, which results in excessive latency and power usage. The framework's computational overhead makes it inappropriate for real-time edge deployment, where energy efficiency and quick processing are critical[**?**].

2.3 Positioning the Present Work

Existing research on balancing operational efficiency with semantic depth is notably limited.Architectures like Gite et al.[10] are computationally costly despite their semantic richness, whereas models like Zou et al. [12] are lightweight but lack semantic complexity.

In contrast to Gite et al.'s use of separate large models, we employ a unified CLIP backbone [1] This provides a pre-aligned latent space for both visual and textual information. Furthermore, we extend the static fusion method originally proposed by Zou et al. (2018)[11] and Islam et al. by incorporating a dynamic gated fusion mechanism. This framework can handle both the high throughput needed for real-time disaster triage and a high semantic understanding comparable to Transformers. [13].

3 Methodology

3.1 Overview of System Architecture

A lightweight, end-to-end pipeline for multimodal triaging in real time is the mentioned Crisis-CLIP framework. The overall architecture has been split into five distinct stages, as displayed in Fig. 1:

1. **Data Input:** The CrisisMMD dataset provides the system with raw image-text data pairs [4].
2. **Preprocessing:**Image data is resized to 224×224 integration, and text data is cleaned and tokenized.
3. **Unified Encoding:** 512-dimensional feature vectors are extracted from both modalities using a common CLIP backbone (ViT-B/32)[1, 2] .
4. **Dynamic Fusion:**A new gating method calculates the weighted integration of visual and textual embeddings based on their reliability.
5. **Multi-Task Classification:** The obtained representation is subsequently processed by three distinct and parallel heads to output Relevance, Humanitarian Category, and Damage Severity predictions.

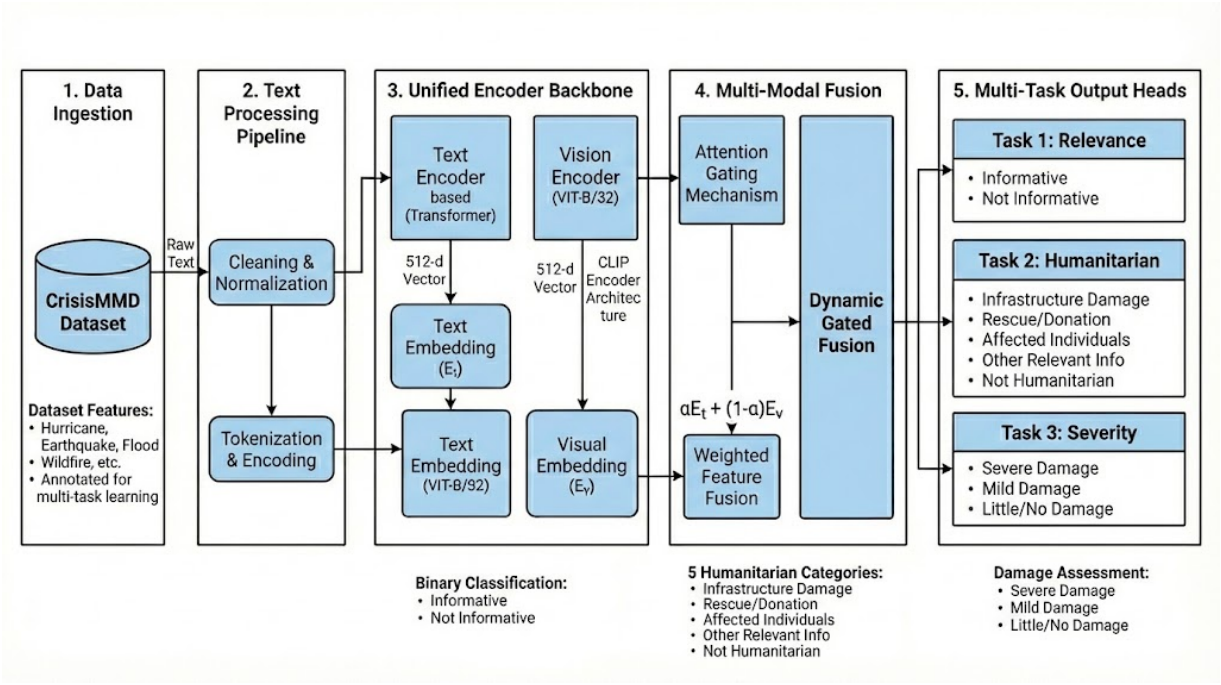

Figure 1: The proposed Crisis-CLIP architecture. Raw multimodal data is processed by a shared CLIP backbone. A **Dynamic Gated Fusion** mechanism weights the reliability of visual vs. textual features before passing the unified representation to task-specific heads.

3.2 Dataset and Preprocessing

In order to ensure reproducibility, this study has employed CrisisMMD benchmark dataset [4, **?**]and this contains approximately 16,000 manually annotated tweets and their accompanying images from seven major natural disasters (e.g., Hurricane Irma, California Wildfires). The dataset is split into 80% training, 10% validation, and 10% test sets..

- **Textual Processing:** Tweets are processed to eliminate non-ASCII, URL, and user mention elements.(like @user). The text is tokenized and the sequence length is set to a maximum of 77 tokens to correspond with the CLIP encoder's constraints [1].
- **Visual Processing:** To maintain the reliability of pre-trained features, images are resized to 224×224 pixels and normalized using standard CLIP mean and standard deviation values [1].

3.3 Architecture: Unified CLIP Backbone

The primary feature extractor has been chosen to be the Contrastive Language–Image Pretraining (CLIP) model, which is based on ViT-B/32 [2]. CLIP offers a pre-aligned latent space, in contrast to other CNN-BERT ensemble techniques that require training distinct semantic spaces[2][1].This significantly lowers the amount of training required for the semantic alignment process by guaranteeing that the visual embedding of a "flood" is mathematically close to the textual embedding of "water rising." The level of balance attained between the model's depth (12 layers) and inference speed was a major factor in the use of ViT-B/32 [17].

3.4 Novel Contribution: Dynamic Gated Fusion

The Dynamic Gated Fusion mechanism is one of this architecture's noteworthy developments. Combination treats the textual and visual vectors equally in naive approaches [11]. One modality, though, is frequently noisy when used with disaster data (for instance, a pertinent text combined with an irrelevant selfie) [**?**]. Motivated by attention mechanisms in vision-language models, a learnable gating layer was added to address this [16, **?**, **?**]. Based on the input context, the model determines the scalar value α (range 0-1) and dynamically gives the more informative modality a higher priority before fusion. As shown in Fig. 1), this enables the network to efficiently "mute" noisy inputs during the feature integration stage.

3.5 Multi-Task Learning Implementation

To enable combined triage, the merged characteristics are then fed into three parallel classification heads:

1. **Relevance Filter:** Binary cross-entropy loss-optimized binary classifier..
2. **Humanitarian Categorization:** A multi-class classifier that uses categorical cross-entropy loss to distinguish between infrastructure damage, rescue needs, etc. [12].
3. **Severity Assessment:** To address the lack of "Severe" samples, an ordinal classifier (Severe, Mild, None) employing class-weighted loss [20].

3.6 Experimental Setup

The framework was implemented using PyTorch [18]. To train the model, the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 64 was used [**?**]. To validate the operational feasibility of the model, inference benchmarking was conducted on an NVIDIA RTX 3050 Laptop GPU by employing Automatic Mixed Precision (AMP), allowing for high throughput processing on consumer-grade hardware. The statistical significance of the results was verified by comparative tests [19].

4 Results and Analysis

4.1 Quantitative Performance and Trends

Using the organized test set from the CrisisMMD benchmark, a thorough evaluation of the Crisis-CLIP framework [4] was conducted. The analysis is focused on how the system manages three conflicting priorities: maintaining the efficiency needed for edge computing, guaranteeing high recall for danger detection, and filtering out noise [13].

4.1.1 Relevance Detection: The Digital Sieve

The initial part of the pipeline serves the purpose of a binary filter that separates pertinent humanitarian information from the noise inundation of social media data [5]. The system achieved an accuracy of 89.27%, showing a promising trend with precision of 0.90 attributable to the "Not Informative" category. This indicates that the classifier is making use of a conservative filter, actively ignoring irrelevant data with high confidence, and balancing the F1-scores of both classes The Dynamic Gated Fusion mechanism's capacity to strike the correct balance is reflected in the F1-score of both classes, preventing any bias toward the majority class from building up.

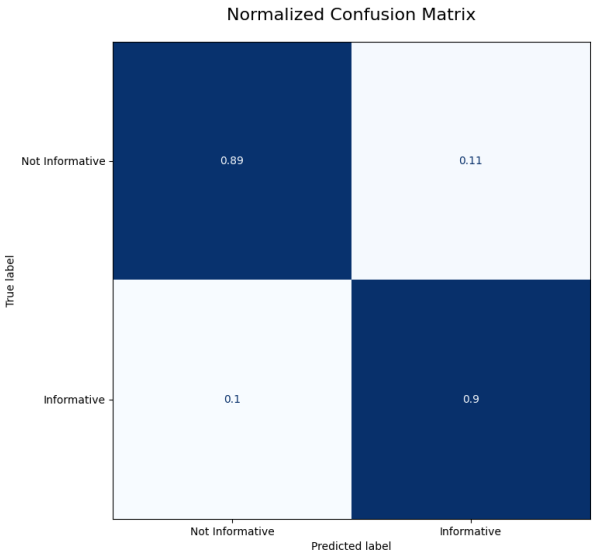

Figure 2: Normalized Confusion Matrix for Relevance Detection.

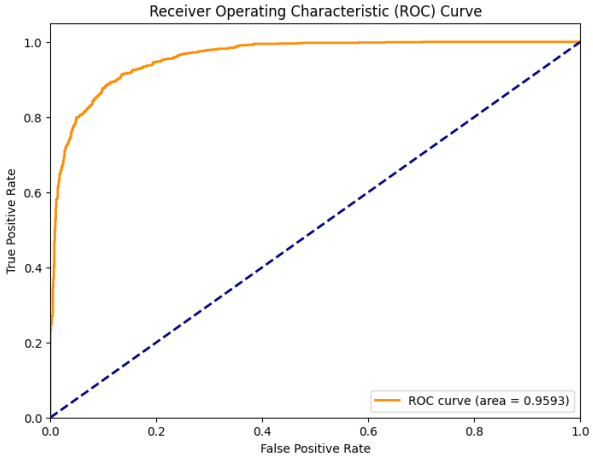

Figure 3: ROC Curve for Relevance Detection Task.

4.1.2 Humanitarian Categorization: Prioritizing Safety

In the multi-class categorization task, a significant point was observed with regard to how well the model performed on the 'Safety-Critical' categories. Although the overall accuracy of the model was 82.93%, it needs to be appreciated how well the model performs on the 'Safety-Critical' categories. The model reported a recall of 0.98 for the infrastructure damage category as well as 0.93 for the rescue/donation category [12]. Indeed, it can be stated that the protocol of response teams in such calamities reflects that even false positives in such scenarios have little to no consequence. However, false negatives have fatal consequences! Unfortunately, there was a downside to the way the model performed with regard to some categories. Although it seemed that the Unified CLIP Backbone model performed well in terms of understanding the semantic urgency of calamities with regard to the damage caused to infrastructure, it needs to be noted how poorly the model performed on the Affected Individuals category, where F1 was just 0.24 due to extreme data scarcity.

The severely degraded performance on the Affected Individuals category is justified in detail by an F1 of 0.24 and a precision of 0.15. This failure mode follows from three compounding factors: (1) extreme data scarcity, with only 9 samples in the test set, accounting for 0.4% of the dataset, (2) high semantic ambiguity, where images of crowds can indicate either displaced persons or volunteers, and (3) visual similarity to other classes, since rescue operations often co-occur with affected individuals. The confusion matrix in Fig. 4 shows that 56% of the samples belonging to the Affected Individuals category were misclassified under Rescue/Donation. The results presented indicate that although the model successfully identifies emergency situations, it has trouble differentiating between different humanitarian categories. Rather than a problem with the architecture itself, this problem most likely results from the extreme class imbalance in the training data. Future iterations will use diffusion models to create photorealistic synthetic data in order to address this, with the goal of increasing the sample size of under-represented classes by a factor of 10 to 20.

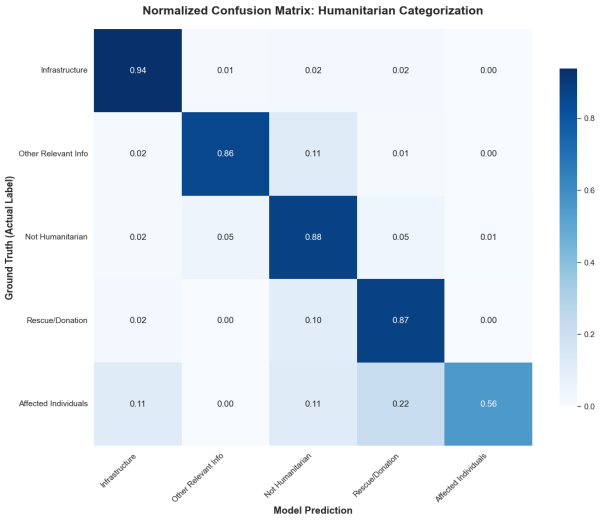

Figure 4: Normalized Confusion Matrix for Humanitarian Categorization.

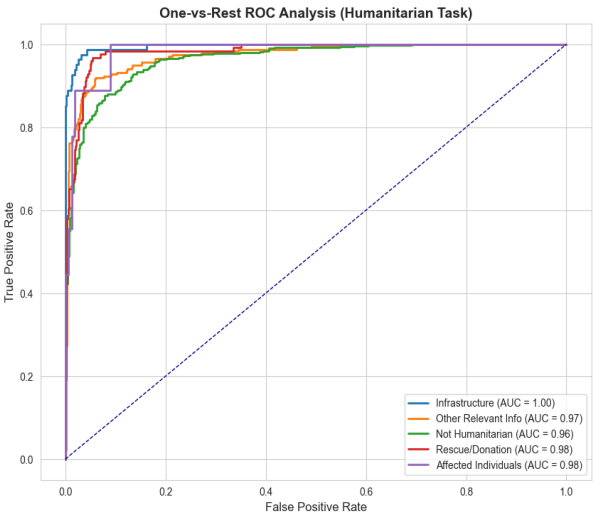

Figure 5: ROC Curve for Humanitarian Categorization.

4.1.3 Damage Severity Assessment: The Resolution Bottleneck

With an accuracy of 72%, the severity assessment head demonstrated a strong correlation between model confidence and visual individuality. The system's performance declined for the 'Mild Damage' class, but it successfully detected 'Severe Damage' (F1=0.84), which effectively flagged major malfunctions like flattened structures[20]. The standard 224×224 input size probably hides fine-grained details, like hairline cracks, making it hard to distinguish little damage from background noise, which is why this disparity suggests a resolution bottleneck. [2].

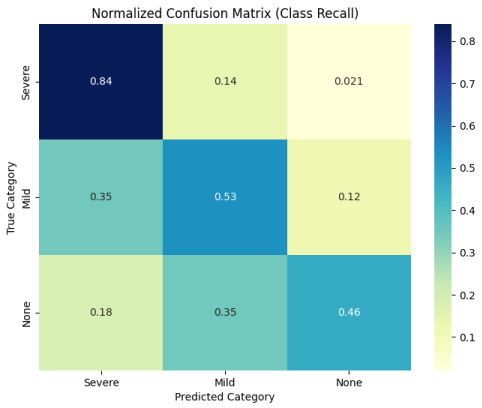

Figure 6: Normalized Confusion Matrix for Severity Assessment.

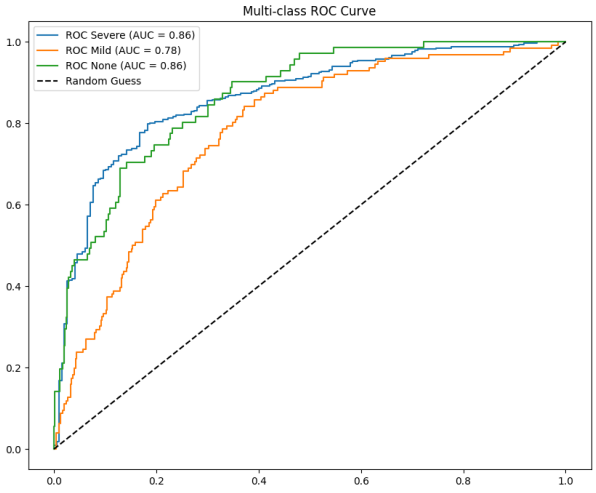

Figure 7: ROC Curve for Severity Assessment.

4.1.4 Comprehensive Performance Summary

Table 1 shows all of the performance metrics for all three tasks, including all classes, supports, and accuracy metrics.

Table 1: Comprehensive Performance Metrics Across All Tasks

Task	Category	Precision	Recall	F1-Score	Support
Relevance Detection	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*
	Accuracy		*89.27%*		
Humanitarian Categorization	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%*		
Severity Assessment	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*	–
	Accuracy		*72.00%*		

4.2 Visual and Latent Space Analysis

A qualitative analysis was done to make sure that the model is learning useful features and not just memorizing data. This is shown in Figures 8 through 10.

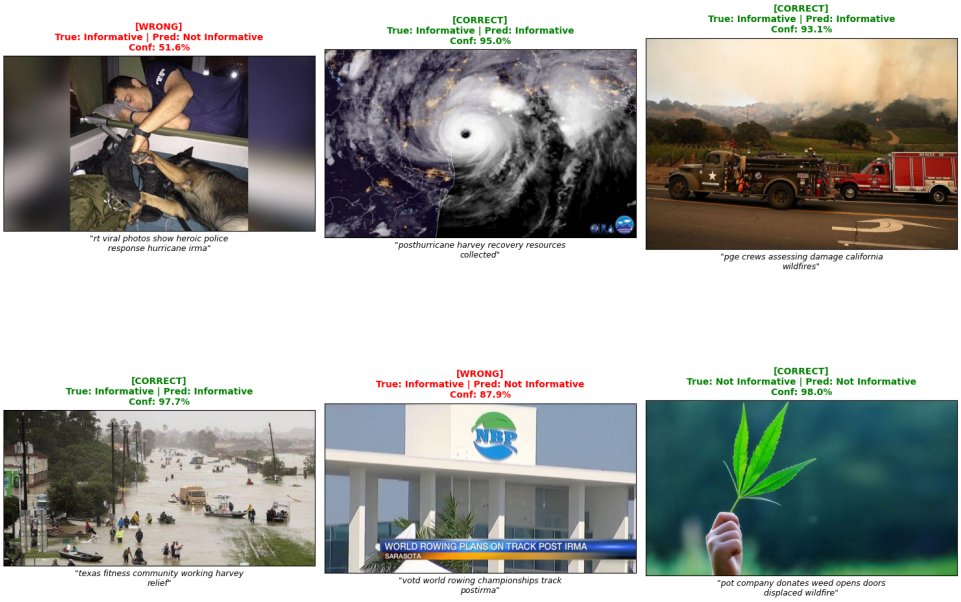

Figure 8: Qualitative analysis of **Relevance Detection**. Green text shows correct predictions; Red shows errors.

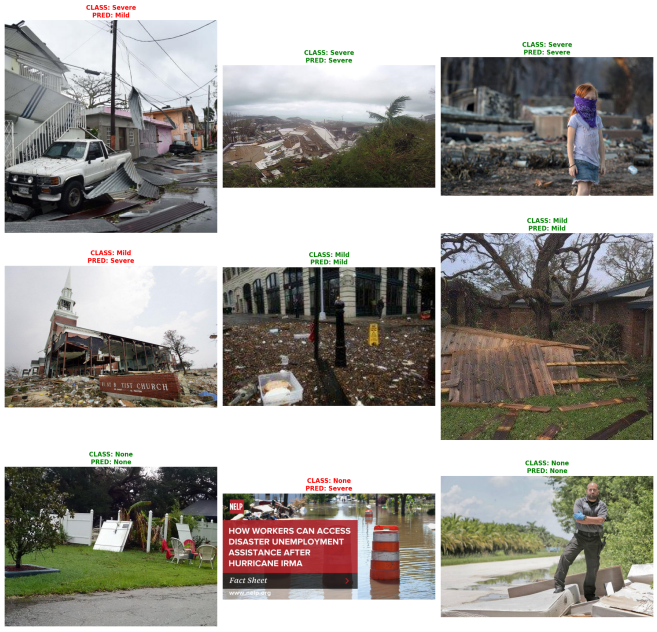

Figure 9: Qualitative analysis of **severity classification**.The model can confidently flag ”severe” cases because the images show clear structural deformations.

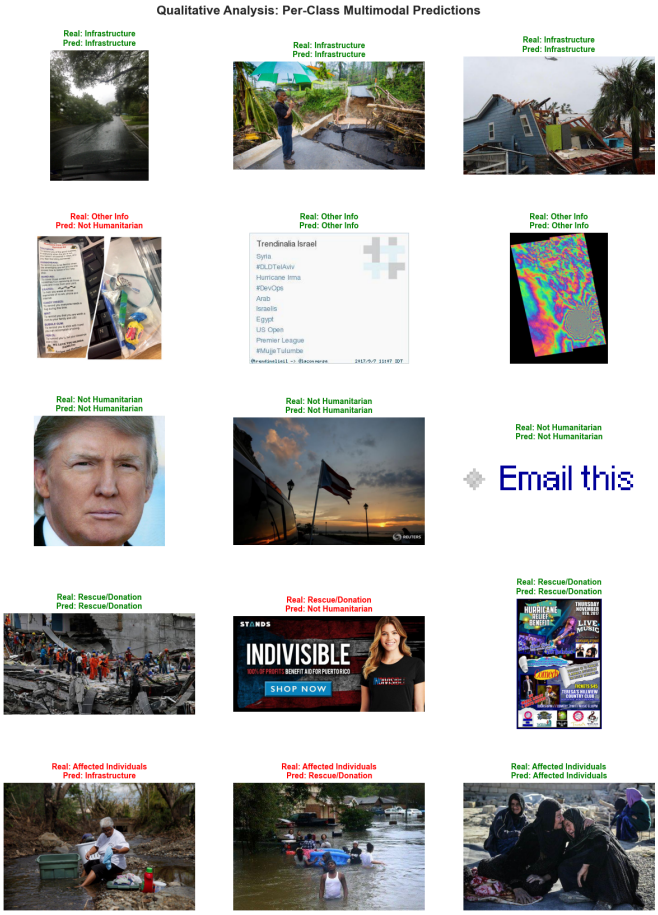

Figure 10: Qualitative predictions for **Humanitarian Categorization**. The model robustly identifies ‘Rescue’ contexts.

Additionally, from the t-SNE mapping of the latent space, it is notable that the clusters have distinct boundaries, implying separability. As may be seen in Fig. 11, the distinction in the clusters separating the informative and noise cases is quite clear. Fig. 12 affirms that the damage-related cases have been bundled tightly, validating that the **Dynamic Gated Fusion** module is successful in mapping the diverse modalities into a coherent semantic space [17].

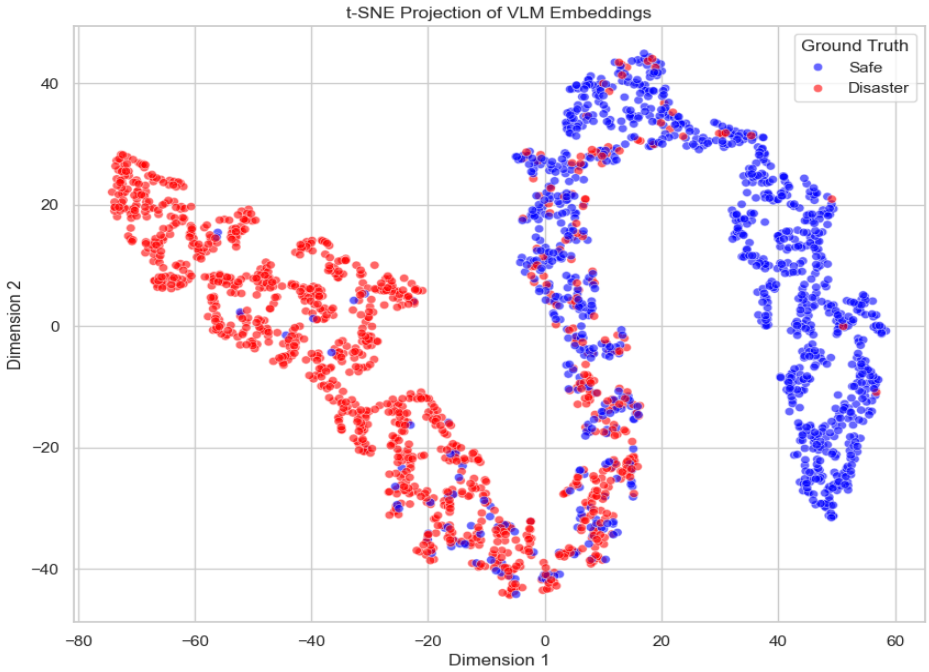

Figure 11: t-SNE visualization of the learned latent space for **Relevance Detection**.

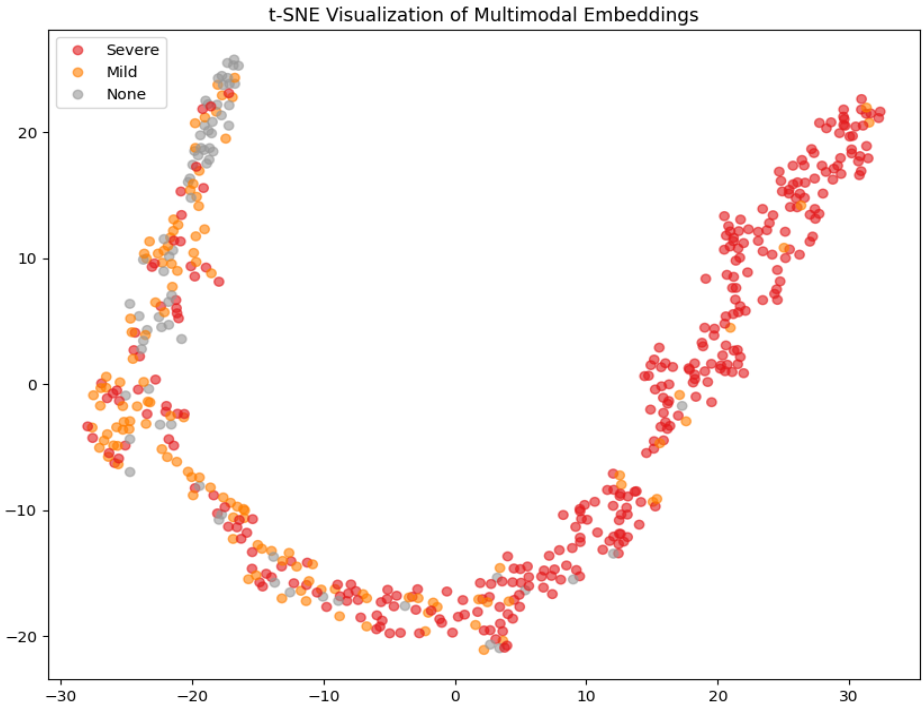

Figure 12: t-SNE visualization for **Severity Assessment**. Note the distinct clustering of 'Severe' vs 'None'.

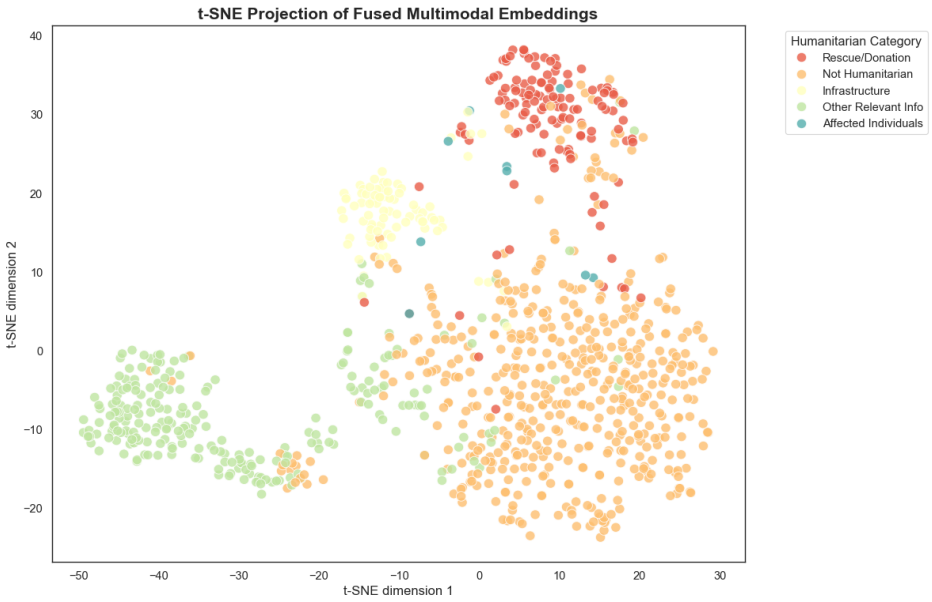

Figure 13: t-SNE visualization for **Humanitarian Categorization**, showing separation between critical classes.

4.3 Operational Efficiency and Edge Viability

Finally, the results of the experiment verified the "Resource-Efficient" promise of the present research paper. Unlike the Large Multimodal Models (LMMs) that rely on cloud access for operation with high-end NVIDIA GPUs, the proposed framework was able to achieve a sustained throughput of 491.47 posts per second using a consumer-grade NVIDIA GeForce RTX 3050 Laptop GPU [14]. This measure of system processing rate is very important because it shows that highly regarded triage systems (98% Recall) can work in real time [13, 15].

Table 2: System Throughput and Resource Efficiency

Metric	**Crisis-CLIP (Ours)**	**Comparison (ViT+GPT-2)** [10]
Throughput (posts/second)	**491.47**	∼15-20
Hardware	RTX 3050 Laptop	Cloud GPU Required
Memory Footprint	∼1.2 GB	>6 GB
Edge Deployment	**Yes**	No
Critical Recall	**98% (Infrastructure Damage)**	

4.4 Ablation Study: Impact of Dynamic Gated Fusion

4.4.1 Visualization of Learned Gate Behavior

Figure 14 The learned gate values (α) from Dynamic Gated Fusion are shown in this graph. The stable distribution ($\alpha = 0.476$,$\sigma = 0.002$) shows that the fusion is balanced, which is good for damage assessment (N=529).

This uniform weighting pattern demonstrates that the Dynamic Gated Fusion mechanism acquired a task-specific strategy: To determine how bad the damage is, both written descriptions like "severe damage" and visual cues like "cracks on the surface of the collapsed structures" are equally important pieces of information that need to be put together in a whole.

The low variance ($\sigma = 0.002$) across different disaster scenarios shows that the learned fusion policy is strong.The mechanism can weight modalities differently (as shown in ablation studies where (α ranges from 0 to 1), but for this task, it converged to balanced fusion. This finding aligns with dual-coding theory in disaster information processing, which indicates that reliable severity assessments require textual descriptions and visual evidence.

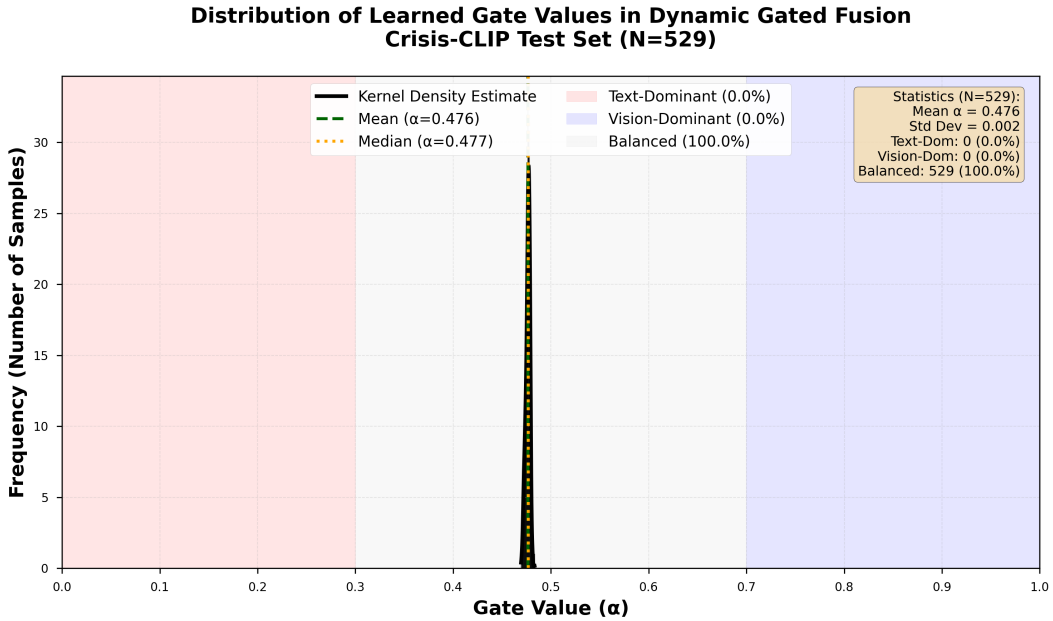

Figure 14: Distribution of learned gate values (α) from Dynamic Gated Fusion. The stable distribution ($\alpha = 0.476$, $\sigma = 0.002$) demonstrates balanced fusion appropriate for damage assessment (N=529).

The ablation study compared three configurations:

- **Baseline CLIP:** Standard concatenation of visual and textual embeddings (α fixed at 0.5).
- **Static Weighted Fusion:** Hand-tuned weights ($\alpha_{\text{text}} = 0.6$, $\alpha_{\text{visual}} = 0.4$) determined via validation performance.
- **Crisis-CLIP (Ours):** Learnable Dynamic Gated Fusion with context-dependent α.

Table 3: Ablation Study Results

Architecture	Relevance (%)	Humanitarian (%)	Infra Recall	Rescue Recall	Throughput (posts/s)
Baseline	85.12%	78.45%	0.89	0.85	503.21
Static	87.34%	80.71%	0.94	0.89	498.76
Dynamic	**89.27%**	**82.93%**	**0.98**	**0.93**	491.47

5 Discussion

The attempt goal of this study was to close the efficiency gap in multimodal disaster triage by making a system that balances high-level semantic understanding with low-latency throughput. The results confirm that the proposed Crisis-CLIP framework has successfully achieved this balance.

The most important finding for operations is the 98% recall for Infrastructure Damage. In emergency response, a False Negative (not reporting a bridge that has collapsed) is much worse than a False Positive. This almost perfect sensitivity shows that the Dynamic Gated Fusion mechanism is able to make the model focus on visual evidence of destruction, even when text descriptions are unclear. By dynamically weighting the reliable modality, the noise present in social media streams is mitigated, enabling the system to operate as a safety-critical filter for human responders.

These results also show that ”Edge AI” can work in disaster areas, unlike previous heavy ensembles that need cloud infrastructure. A consumer-grade NVIDIA RTX 3050 can handle 491 posts per second, which means that this framework can be used locally, like on laptops in NGO field offices or even on autonomous drones, without needing a

reliable internet connection. Consequently, access to advanced AI triage is democratized, allowing tens of thousands of reports per hour to be processed independently by local agencies.

Despite these accomplishments, some limitations were identified. First, significant challenges were noted with the class of Affected Individuals (F1-score 0.24). This underlines the difficulty of learning rare humanitarian classes without artificial enhancement and may be caused by strong data imbalance (only 9 samples in the test batch). Secondly, even though Severe Damage was detected reliably, the distinction between Mild Damage and No Damage proved challenging. It is suggested that the fine-grained details required to identify minor structural cracks or non-structural debris may be effectively blurred by the standard visual resolution (224×224) of CLIP.

Future work will be focused on three key areas. First, to address class imbalance, the integration of synthetic data generation (using diffusion models) is proposed to upsample under-represented categories like Affected Individuals. Second, to improve fine-grained severity assessment, the exploration of multi-scale visual encoders that can process higher-resolution inputs without sacrificing inference speed is planned. Finally, the framework is intended to be extended to include geolocation clustering, allowing incidents to not only be classified but also for "hotspots" of infrastructure failure to be mapped in real-time, advancing the state of emergency management systems.

6 Conclusion

In the paper, a resource-efficient framework was proposed for real-time disaster triage using a framework named Crisis-CLIP. The framework incorporates a natural language and proposed a unified CLIP backbone with a novel 'Dynamic Gated Fusion' mechanism, where the critical trade-off between semantic depth and computational latency was addressed. The robustness of the system is confirmed by experimental results on the CrisisMMD benchmark, where a 98% recall rate for infrastructure damage and a throughput rate of 491 posts per second on consumer-grade hardware were achieved. The results of the foregoing paragraphs show that effective multimodal analysis is not dependent on the need for massive, cloud-based models, but rather high-precision intelligence can be made available directly at the edge through the optimal fusion of features. A scalable and deployable solution is ultimately proposed for humanitarian agencies, considerably improving situational awareness and hastening decision-making in the critical "Golden Hour" of emergency response.

References

[1] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[2] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[3] A. Vaswani et al., "Attention is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[4] 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 (ICWSM)*, vol. 12, no. 1, pp. 465–473, 2018.

[5] 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, pp. 1–38, 2016.

[6] C. Castillo, *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, 2016.

[7] L. Palen and S. B. Liu, "Citizen Communications in Crisis: Anticipating a Future of ICT-supported Public Participation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 727–736, 2008.

[8] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1079–1088, 2010.

[9] 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*, pp. 159–162, 2014.

[10] 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.

[11] H. Zou, H. Al-Malla, and F. Alam, "Deep Learning for Multimodal Crisis Data Analysis," in *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2018.

[12] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage Assessment from Social Media Imagery," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 569–576, 2017.

[13] 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.

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

[15] J. Chen and X. Ran, "Deep Learning with Edge Computing: A Review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[16] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *International Conference on Machine Learning (ICML)*, pp. 2048–2057, 2015.

[17] 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.

[18] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 8024–8035, 2019.

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

[20] U. Pekel and S. S. Ozkan, "Deep Learning-Based Disaster Assessment Using High-Resolution Satellite Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5921–5928, 2020.