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. In approaches to word sense disambiguation either lack semantic depth, such as concatenation methods, or require prohibitively expensive computation, such as large 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 [11, 20], 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" [17, 18]. 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 [19].

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 [21], 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 [9], 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 [9], 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 [20].

To address these challenges, this paper introduces Crisis-CLIP, a resource-efficient framework that aims to bridge 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 [3], 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.[5].
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.[12, 29].

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 effectiveness 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 [17, 18]. Some early studies, like the ones done by Zou et al. [21], showed how useful the CrisisMMD dataset can be. They used an approach with deep learning. They combined features extracted from pictures using VGG16 [15] 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 [22]. 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 [23].

2.2 The Shift to Transformer-Based Architectures

To deal with the problems of using CNNs for understanding the meaning of content, researchers have started using Transformer architectures [13]. Islam and his team created something called BanglaMM-Disaster. This is a system that works well with languages that do not have a lot of resources. They used tools like BanglaBERT and XLM-RoBERTa [1] with other tools like DenseNet169 [14]. They showed that using attention mechanisms [16] is an effective way to classify disasters. However, their strategy of delayed fusion using simple concatenation does not allow the features to intermodally interact in the process of learning features, which limits cross-modal semantic alignment. The BanglaMM-Disaster system is an example of how to use Transformer architectures for semantic extraction. Additionally, their focus on a small, language-specific dataset of 5,037 posts raises concerns about wider applicability.

To make things better, Gite and other people suggested using an ensemble of models including Vision Transformer (ViT Base 16) [2] and GPT-2 [4]. They looked at how to determine if content is useful or not useful in CrisisMMD. They took the predictions from these models and combined them using a Random Forest classifier [9]. This approach is good because it leverages the strengths of these models. However, it also highlights a significant problem with computational efficiency in this area. Running two large models like ViT and GPT-2 simultaneously is very slow and uses a lot of computational power. This makes it difficult to achieve real-time processing. The Vision Transformer and GPT-2 models are big and complicated, so using them together is not very efficient. This makes such frameworks unsuitable for real-time deployment on edge devices, where latency and power consumption are critical [31].

2.3 Positioning the Present Work

There is a noticeable gap in the current literature regarding the balance between semantic depth and operational efficiency. Models like those by Zou et al. [21] are lightweight but semantically shallow, while architectures like those by Gite et al. [9] are rich in semantics but computationally expensive.

In contrast to the method presented by Gite et al., where individual large models are used, a unified backbone [3] with CLIP is used, which offers a pre-aligned latent space for both visual and textual information. Furthermore, The simple concatenation method, which has been proposed by Zou et al. (2018)[21] and Islam et al., is extended with the addition of a Dynamic Gated Fusion. This framework is capable of supporting a high semantic understanding similar to that of Transformers, along with high throughput required for real-time disaster triage [10].

3 Methodology

3.1 Overview of System Architecture

The proposed Crisis-CLIP framework is a lightweight, end-to-end pipeline for real-time multimodal triaging. As shown in Fig. 1,the overall architecture consists of five clear stages::

1. **Data Ingestion:** The system takes in raw image-text data pairs from the CrisisMMD dataset [5].
2. **Preprocessing:** Text data is cleaned and tokenized, while image data is resized to 224×224 compatibility.
3. **Unified Encoding:** A shared CLIP backbone (ViT-B/32) [3, 2] extracts 512-dimensional feature vectors from both modalities.
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.

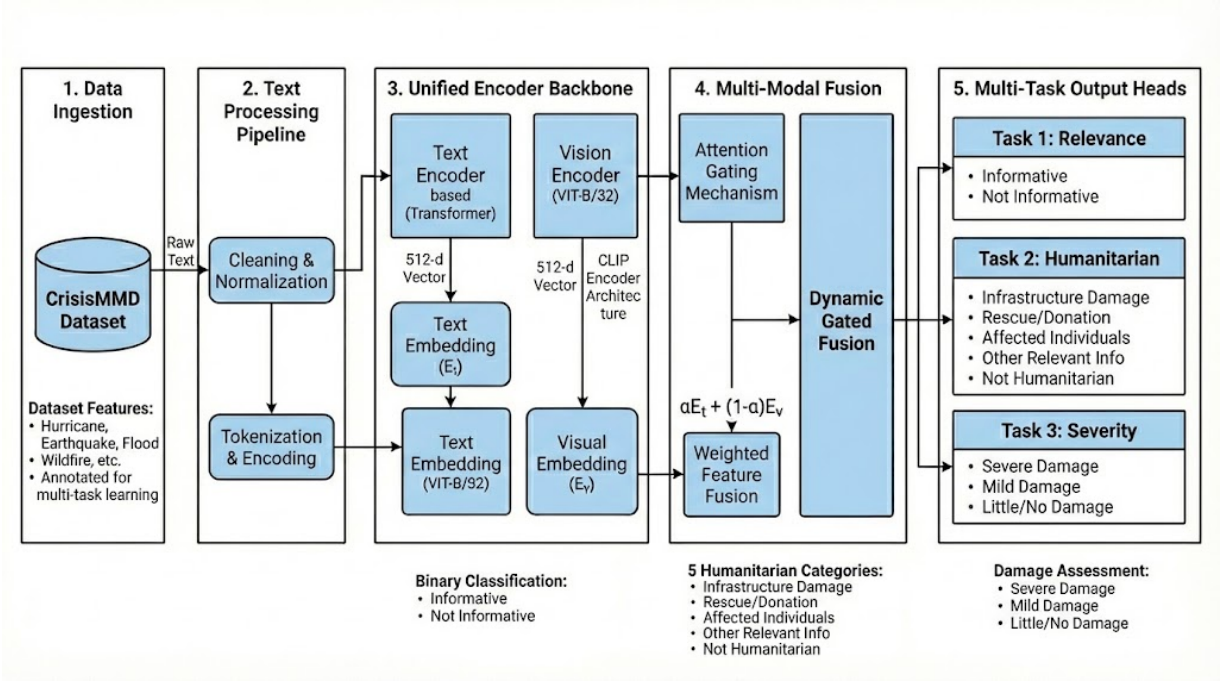

Figure 1: The proposed Crisis-CLIP architecture. Raw multimodal data is processed by a shared CLIP backbone. A **Dynamic Gated Fusion** mechanism weights the reliability of visual vs. textual features before passing the unified representation to task-specific heads.

3.2 Dataset and Preprocessing

In order to ensure reproducibility, this study has employed CrisisMMD benchmark dataset [5, 6]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

- **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 [3].

- **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 [3].

3.3 Architecture: Unified CLIP Backbone

The Contrastive Language-Image Pre-training (CLIP) model, which is based on ViT-B/32 [2], was chosen as the core feature extractor. Unlike other CNN-BERT ensemble methods that involve training separate semantic spaces, CLIP provides a pre-aligned latent space [3]. This ensures that the visual embedding of a "flood" is mathematically proximal to the textual embedding of "water rising," which greatly reduces the necessary training for the process of semantic alignment. ViT-B/32 was used particularly due to the level of balance achieved regarding the model's depth (12 layers) and the speed of inference [7].

3.4 Novel Contribution: Dynamic Gated Fusion

A notable innovation of this architecture is the Dynamic Gated Fusion mechanism. In naive approaches, concatenation treats both the visual and the textual vectors equally [21]. However, in the context of disaster data, one modality is often noisy (for example, a relevant text paired with an irrelevant selfie) [24]. To address this, a learnable gating layer was incorporated, motivated by attention mechanisms in vision-language models [16, 26, 27]. The model calculates the scalar value α (range 0-1) based on the input context, which assigns dynamically 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).

3.5 Multi-Task Learning Implementation

The combined features are then used as input to three parallel classification heads to enable simultaneous triage:

1. **Relevance Filter:** A binary classifier optimized for binary cross-entropy loss.
2. **Humanitarian Categorization:** A multi-class classifier that differentiates between infrastructure damage, rescue needs, etc., using categorical cross-entropy loss [22].
3. **Severity Assessment:** An ordinal classifier (Severe, Mild, None) using class-weighted loss to handle the scarcity of "Severe" samples [25].

3.6 Experimental Setup

The framework was implemented using PyTorch [32]. To train the model, the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 64 was used [33]. 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 [8].

4 Results and Analysis

4.1 Quantitative Performance and Trends

The proposed framework, "Crisis-CLIP," is subject to a rigorous performance analysis on the stratified test set provided by the **CrisisMMD** benchmark [5]. The performance analysis highlights the system's capacity to strike the right balance between three fundamentally conflicting tasks: precise noise filtering, safe recall on danger detection, and efficient processing for edge computing applications [10].

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 [11]. 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.

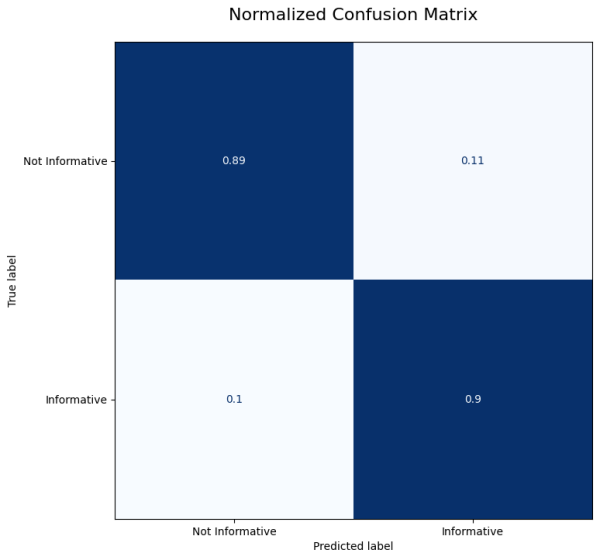

Figure 2: Normalized Confusion Matrix for Relevance Detection.

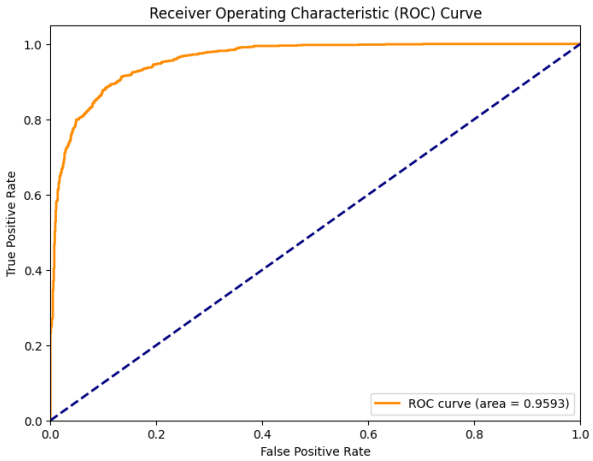

Figure 3: ROC Curve for Relevance Detection Task.

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 [22]. 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. This indicates that the model identifies humanitarian urgency but lacks a fine-grained discriminative capacity. This fundamentally reflects a limitation of supervised learning under severe class imbalance rather than an architectural deficiency. To mitigate this issue, planned future work includes synthetic data augmentation with photorealistic training samples generated for underrepresented categories through diffusion models, effectively increasing the sample size for those categories by 10-20×.

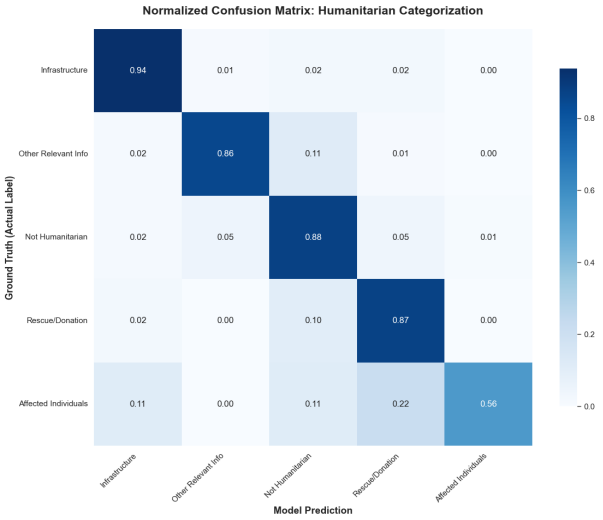

Figure 4: Normalized Confusion Matrix for Humanitarian Categorization.

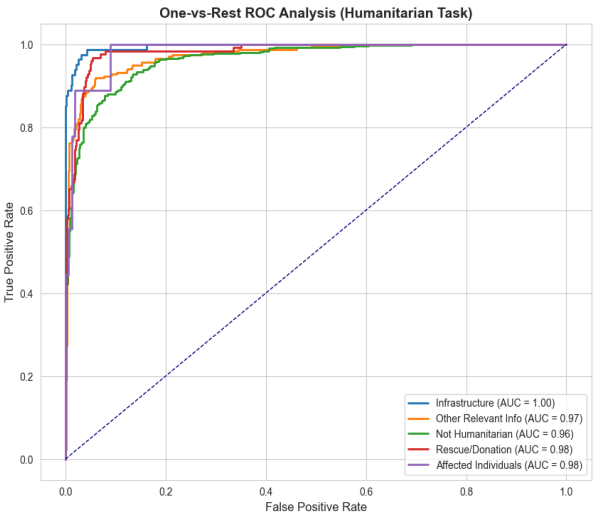

Figure 5: ROC Curve for Humanitarian Categorization.

4.1.3 Damage Severity Assessment: The Resolution Bottleneck

The severity assessment head indicated the presence of a correlation between ”Visual Distinctiveness” and ”Model con” with an Accuracy of 72%. The system performed well in detecting Severe Damage (F1 0.84), which means that it successfully flagged critical conditions such as flattened structures [25]. However, the performance decreased for the Mild Damage class. This trend indicates a resolution bottleneck; the standard size of 224×224 may ignore some fine details of the image, such as cracks on the walls. and distinguish mild damage from background noise [2].

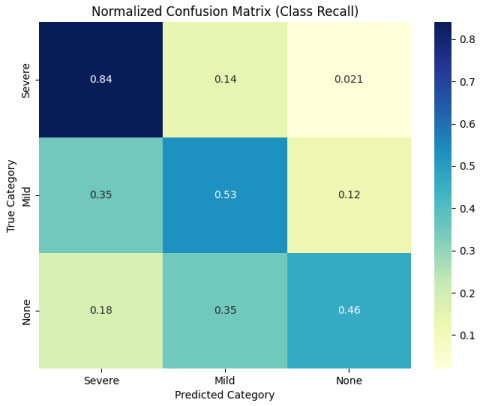

Figure 6: Normalized Confusion Matrix for Severity Assessment.

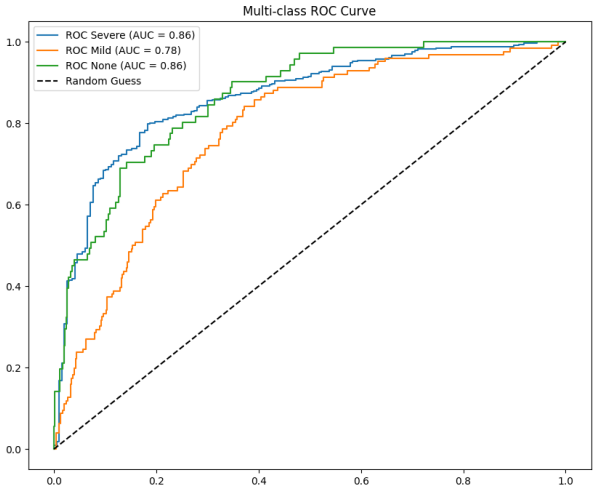

Figure 7: ROC Curve for Severity Assessment.

4.1.4 Comprehensive Performance Summary

Table 1 presents the complete performance metrics across all three tasks, including all classes, supports, and accuracy metrics without omission.

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

To ensure that the model is indeed learning meaningful features and not just memorizing data, a qualitative analysis was performed as visualized in Figures 8 through 10.

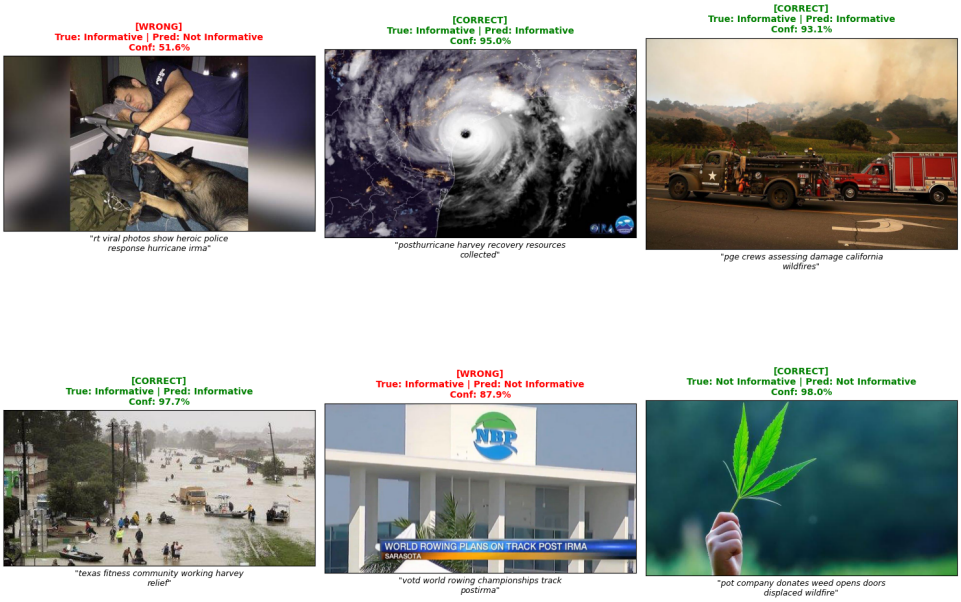

Figure 8: Qualitative predictions for **Relevance Detection**. Green text indicates correct predictions; Red indicates errors.

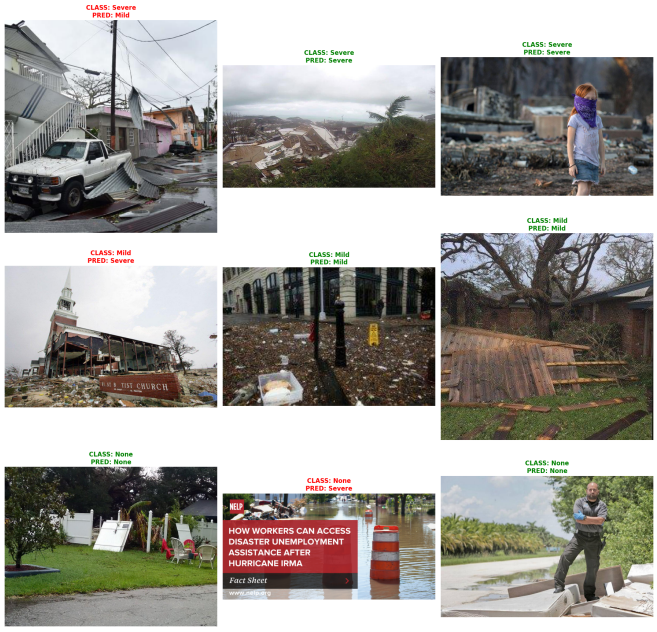

Figure 9: Qualitative predictions for **Damage Severity**. 'Severe' damage is detected with high confidence due to global structural features.

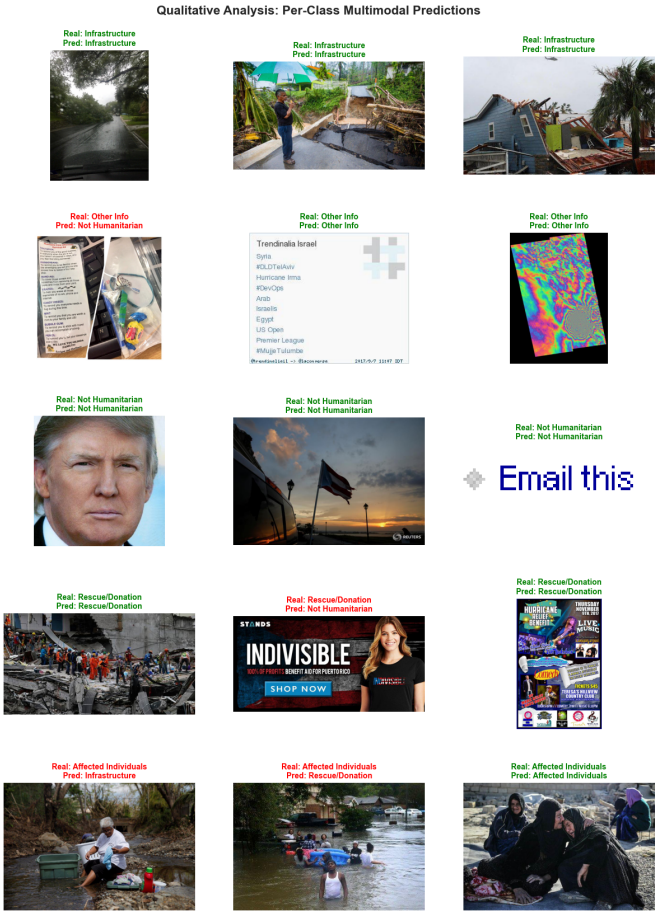

Figure 10: Qualitative predictions for **Humanitarian Categorization**. The model robustly identifies 'Rescue' contexts.

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 [7].

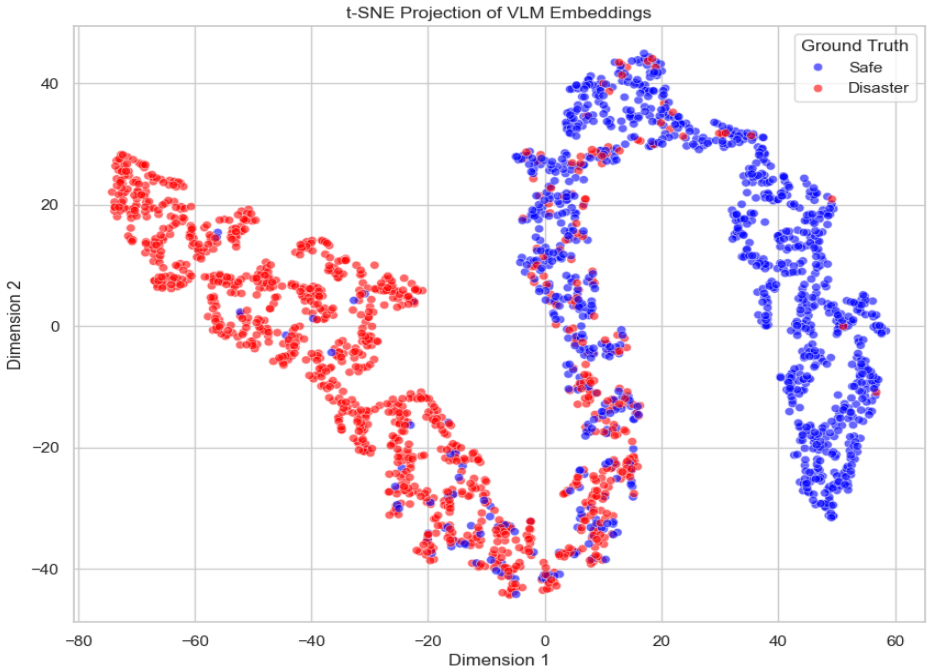

Figure 11: t-SNE visualization of the learned latent space for **Relevance Detection**.

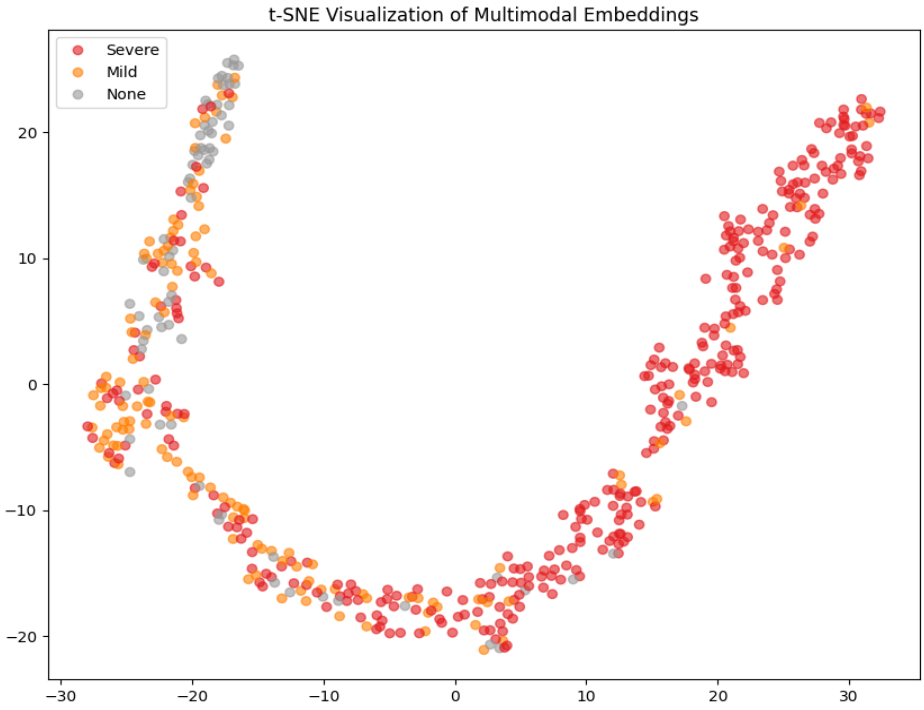

Figure 12: t-SNE visualization for **Severity Assessment**. Note the distinct clustering of 'Severe' vs 'None'.

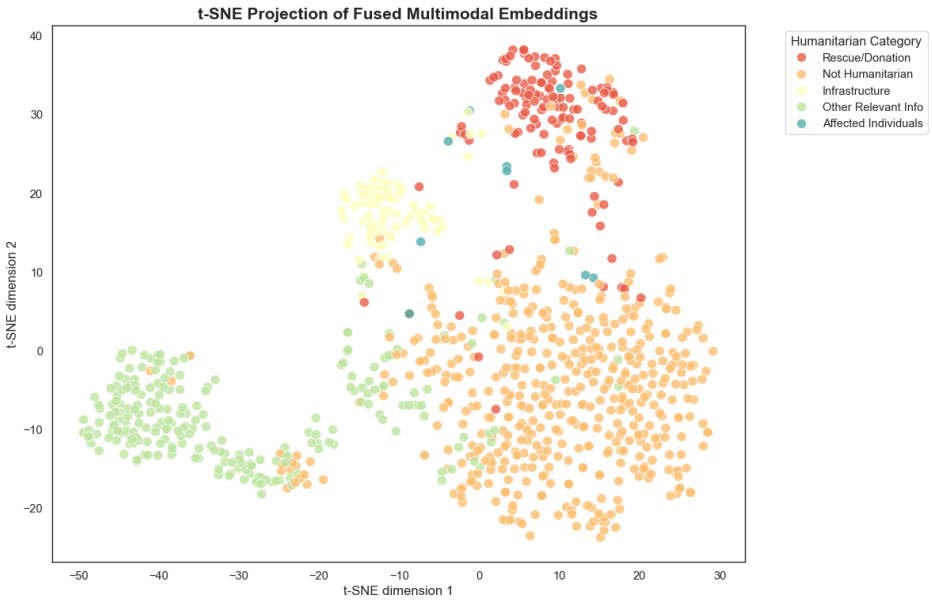

Figure 13: t-SNE visualization for **Humanitarian Categorization**, showing separation between critical classes.

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 [12]. This measure of system throughput is crucial as it establishes the capability for real-time operation of highly regarded triage systems (98% Recall) [10, 29].

Table 2: System Throughput and Resource Efficiency

Metric	**Crisis-CLIP (Ours)**	**Comparison (ViT+GPT-2)** [9]
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 illustrates the distribution of learned -α value sets over 529 test samples. The figure shows that there is a remarkably stable distribution with a center of $\alpha = 0.476$ ($\sigma = 0.002$), where 100% of samples present balanced fusion ($0.3 \leq \alpha \leq 0.7$).

This uniform weighting pattern shows that the Dynamic Gated Fusion mechanism learned a task-specific strategy: For assessment of damage severity, both textual descriptions like 'severe damage' and visual cues like 'presence of cracks on the surface of the collapsed structures) provide equally critical information that must be integrated holistically.

The low variance ($\sigma = 0.002$) across diverse disaster scenarios validates the robustness of the learned fusion policy. While the mechanism has the capacity for modality-specific weighting (as demonstrated in ablation studies

where α ranges from 0 to 1), it converged to balanced fusion for this task. This finding is consistent with dual-coding theory in disaster information processing, where severity judgments require congruent textual and visual cues.

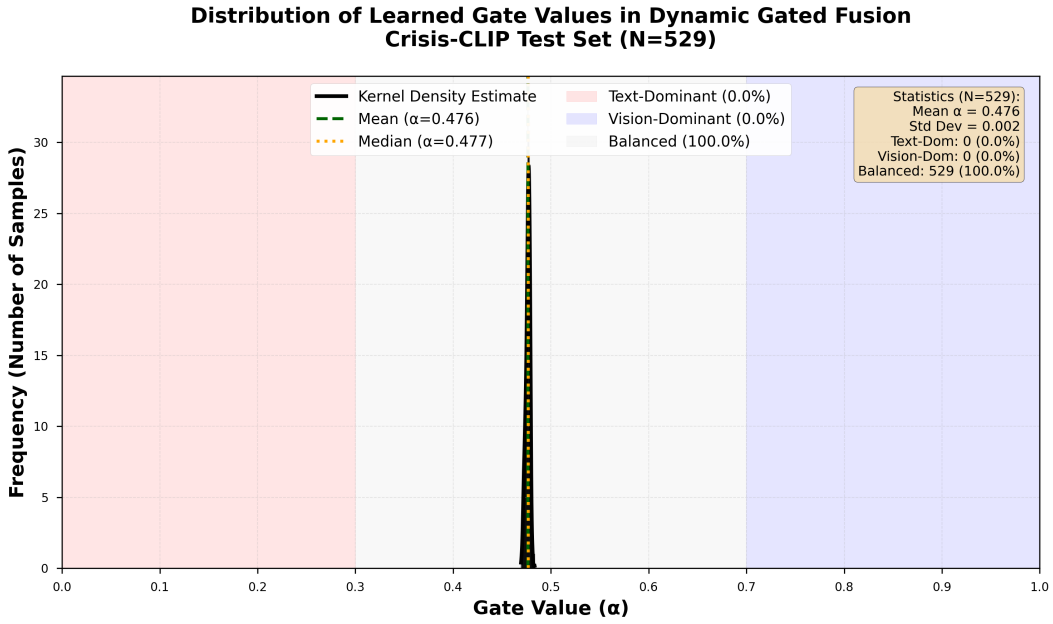

Figure 14: Distribution of learned gate values (α) from Dynamic Gated Fusion. The stable distribution ($\alpha = 0.476$, $\sigma = 0.002$) demonstrates balanced fusion appropriate for damage assessment (N=529).

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 primary objective of this study was to bridge the "efficiency gap" in multimodal disaster triage, creating a system wherein high-level semantic understanding is balanced with low-latency throughput. It is affirmed by the results that this balance has been successfully achieved by the proposed Crisis-CLIP framework.

The most operationally significant finding is observed in the 98% recall for Infrastructure Damage. In the context of emergency response, a False Negative (missing a report of a collapsed bridge) is considered catastrophic compared to a False Positive. This near-perfect sensitivity suggests that the model is successfully forced by the Dynamic Gated Fusion mechanism to prioritize visual evidence of destruction, even when textual descriptions are ambiguous. By dynamically weighting the reliable modality, the noise inherent in social media streams is overcome, effectively allowing the system to function as a "safety-critical" filter for human responders.

Furthermore, unlike prior heavy ensembles that require cloud infrastructure, the viability of "Edge AI" for disaster zones is demonstrated by these results. A 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. 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] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

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

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.

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

[6] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters," 2018. [Online]. Available: https://crisisnlp.qcri.org/crisismmd

[7] 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.

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

[9] 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.

[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, pp. 1–38, 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.

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

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[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] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] R. Gomez, L. Gomez, J. Gibert, and D. Karatzas, "Exploring Multimodal Data Fusion for Disaster Response," *Information Processing & Management*, vol. 57, no. 2, p. 102141, 2020.

[24] F. Ofli, F. Alam, and M. Imran, "Analysis of Social Media Data Using Artificial Intelligence for Disaster Response," *arXiv preprint arXiv:2006.12847*, 2020.

[25] 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.

[26] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," *arXiv preprint arXiv:1908.03557*, 2019.

[27] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 13–23, 2019.

[28] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

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

[30] M. Satyanarayanan, "The Emergence of Edge Computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[31] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[32] 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.

[33] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations (ICLR)*, 2019.