applied sciences

MDPI

Article

LRIF-Net: Decision-Preserving Lightweight Recurrent Interaction Fusion for Multimodal Crisis Information Classification

Ziqi Chen [1,2], Zhian Pan [1,2,*], Shanshan Li [1,2,*], Qingjie Liu [1,2], Huanzhi Gao [1] and Guan Li [3]

[1] School of Computer Science and Information Security, University of Emergency Management, Beijing 101601, China; 25661616@st.cidp.edu.cn (Z.C.); liuqingjie@cidp.edu.cn (Q.L.); gaohuanzhi@cidp.edu.cn (H.G.)
[2] Hebei Province University Smart Emergency Application Technology Research and Development Center, Beijing 101601, China
[3] Planning and Development Department, Big Data Center of Ministry of Emergency Management, Beijing 100054, China; lg1268@126.com
* Correspondence: panzhian@cidp.edu.cn (Z.P.); liss@cidp.edu.cn (S.L.)

Featured Application

Multimodal classification of crisis-related social-media posts to support information triage by emergency analysts.

Abstract

Emergency-information analysts need to organize image-text reports into humanitarian categories before reports can be verified and prioritized. This study asks whether fine-grained evidence refinement can be added to an established classifier without replacing its decision path or greatly enlarging the task-specific update. We propose LRIF-Net, a decision-preserving recurrent residual fusion framework that retains a complete frozen global-local classifier and adds a compact parameter-tied correction branch. The branch is trained and evaluated with three predefined, synthetic evidence-degradation conditions—missing image, missing text, and different-class image mismatch—so the resulting evidence concerns these controlled protocols rather than every form of real-world data corruption. On the canonical 7314-pair CrisisMMD Task 2 split, the fixed three-seed ensemble achieves 91.35% Accuracy, 88.88% Macro F1, and 91.32% Weighted F1. The residual stage updates 517,509 parameters, 88.76% fewer than matched standard cross-attention, while reducing all retained non-CLIP parameters by 59.17%. LRIF-Net provides competitive performance under the matched evaluation protocol, but paired comparisons do not establish predictive superiority and its Balanced Accuracy and Macro F1 are lower than those of the retained direct classifier. Ablations attribute the controlled degradation gains mainly to the learning objective. The study contributes the coordinated design and evaluation of a compact residual refinement mechanism for crisis-information triage, with predictive and robustness claims limited to the reported evidence.

Keywords: crisis information classification; multimodal learning; lightweight fusion; image-text interaction; parameter-efficient adaptation; incomplete multimodal evidence

Check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 2 August 2026
Revised: 5 September 2026
Accepted: 5 September 2026
Published: 8 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

During and after a disaster, large volumes of reports, images, and short messages must be converted into information that emergency managers and humanitarian organizations can review and use. This process involves several groups. Members of the public and field

personnel report events and needs; platform and data teams collect the reports; information-management personnel organize them; and analysts verify their relevance, compare them with other sources, and interpret their possible consequences. Studies of post classification, damage extraction, and resilient edge analytics illustrate how automated analysis can support this broader information process [1–3].

Automated classification belongs near the beginning of this process. Its purpose is to sort incoming posts into preliminary information groups so that relevant material can be reviewed more quickly. It does not replace analysts or make operational decisions. Analysts remain responsible for checking the source and context of a post, resolving ambiguous cases, setting priorities, and deciding whether the information should influence a situation report or response action.

This preliminary screening is difficult because crisis posts are short, context-dependent, and often multimodal. Text may describe a location, need, or response activity, while an image may show damage, affected people, or available supplies. Either modality may be missing, weak, or inconsistent with the other. The class distribution is also uneven, so high aggregate accuracy can conceal poor recognition of less frequent but important categories. A useful classifier must therefore connect local evidence across text and images, retain a dependable decision path when additional interaction is unhelpful, and be evaluated at both aggregate and class levels.

Figure 1 illustrates the five humanitarian information categories used in the evaluated CrisisMMD Task 2 data: affected individuals; infrastructure and utility damage; not humanitarian; other relevant information; and rescue, volunteering, or donation. These labels are defined by the dataset preparation [4,5]. They organize the present closed-set classification experiment, but they are not an exhaustive taxonomy of real-world crisis situations. Posts outside these labels may still be important and require separate analyst review or an open-set recognition procedure.

Previous research addresses different parts of this screening problem. Text-based methods identify humanitarian categories and requests from short messages [6,7]. Multimodal crisis models combine language and visual information using convolutional encoders, attention, gating, graph reasoning, image captions, or pretrained vision-language representations [5,8–15]. These approaches show that the two modalities can provide complementary evidence, but they also differ in the amount and location of cross-modal interaction.

Related work outside crisis classification offers further strategies. Latent bottlenecks and cross-attention support more focused exchange between modalities [16–23]. Parameter-efficient adaptation limits the number of task-specific parameters that must be learned [24–30]. Modality-dropout and missing-modality studies train or test models when part of the evidence is unavailable [31–39]. Together, these studies provide useful components, but they do not by themselves specify how to preserve an established crisis classifier while adding compact, repeated local interaction and evaluating the resulting decisions under class imbalance and controlled evidence loss.

Existing methods leave a more specific question unresolved: how can local evidence be refined without replacing an established classifier or imposing a large adaptation and execution burden? Fine-grained interaction may recover a decisive phrase or image region that global fusion overlooks. However, replacing the original classifier with a new fusion route makes it harder to determine whether a changed prediction comes from the established evidence path or from the added interaction. Performance under synthetically missing or mismatched inputs must also be interpreted only for those conditions; it does not establish robustness to every form of real-world crisis-data degradation.

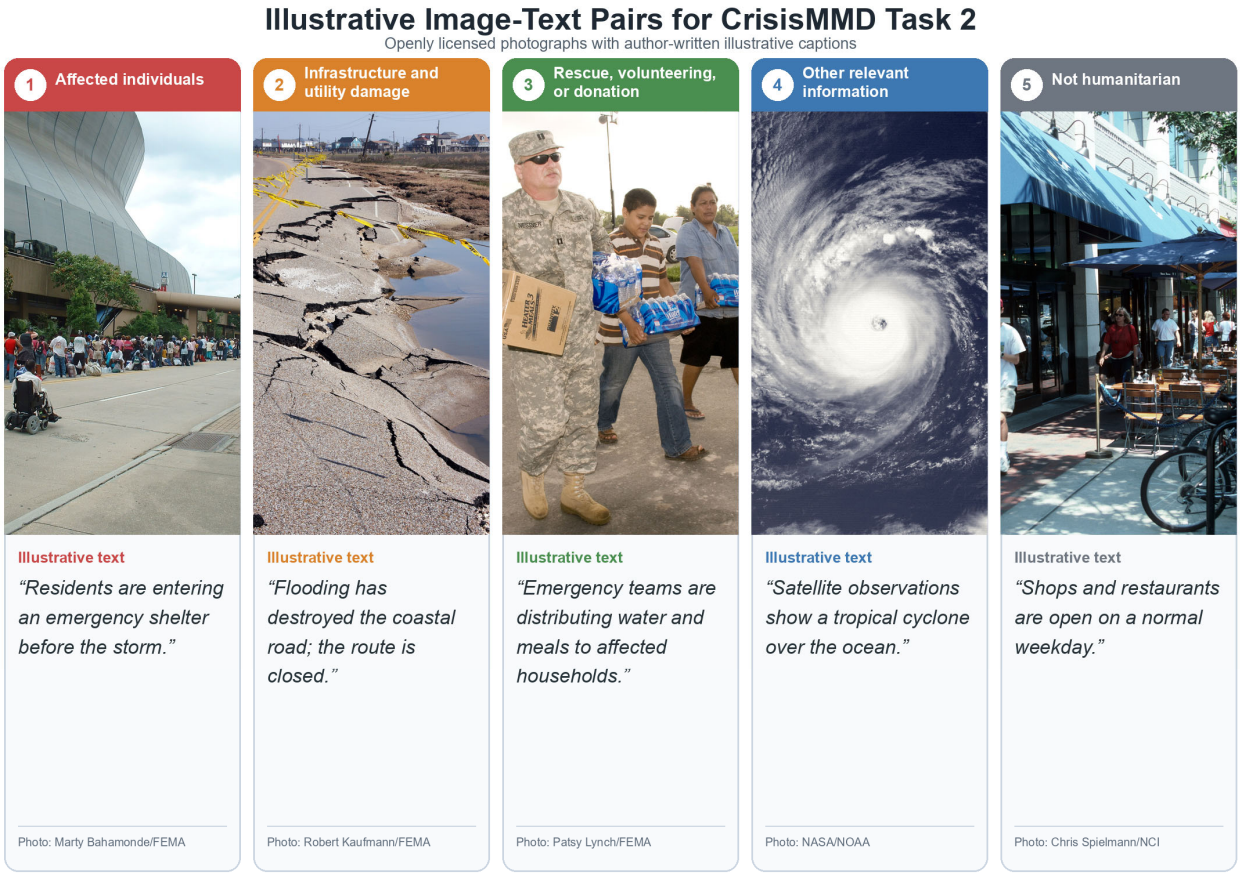


Figure 1. Illustrative image-text examples for the five humanitarian information categories in CrisisMMD Task 2. The photographs are openly licensed and the accompanying captions were written by the authors; these examples are not samples from the CrisisMMD dataset. Photo credits, from left to right: Marty Bahamonde/FEMA; Robert Kaufmann/FEMA; Patsy Lynch/FEMA; NASA/NOAA; and Chris Spielmann/National Cancer Institute.

This study asks whether local image-text evidence can be refined through a compact recurrent residual path while a complete direct crisis classifier remains part of the final decision. LRIF-Net addresses this question by retaining the direct classifier as the prediction basis and learning an additive correction from repeated local interaction. Because the direct logits and the correction remain separately observable, the model can use additional cross-modal evidence without hiding the original decision path. The recurrent interaction reuses its parameters, which limits the task-specific attention budget. Detailed network dimensions, attention equations, loss terms, and training procedures are presented in Sections 3.2–3.9.

The mechanisms used in this design, including latent tokens, cross-attention, parameter sharing, and residual learning, have established precedents. LRIF-Net combines them for a specific purpose in crisis classification. The retained classifier preserves the original decision route, while the parameter-tied recurrent residual branch revisits local evidence without replacing it. Controlled evidence-degradation training and analysis then evaluate how the combined system behaves when specified inputs are missing or mismatched.

The evaluation addresses these questions directly. Clean results and class-level errors measure predictive behavior under imbalance, while matched comparisons and ablations isolate the residual branch's contribution. The controlled missing-image, missing-text, and mismatched-image conditions define the evidence-degradation tests. Event-held-out and input-intervention analyses examine transfer and decision relevance, and parameter and

timing measurements quantify adaptation cost. Conclusions about robustness, generalization, and superiority remain limited to these protocols and the reported statistical evidence.

The main contributions of this work are:

1. A decision-preserving recurrent residual fusion mechanism for crisis-information classification. It keeps a complete global-local classifier active as the direct decision path and constrains a zero-initialized residual branch to provide a separately observable additive correction rather than a replacement prediction.
2. A compact parameter-tied interaction path for iterative token-patch evidence refinement. The same interaction cell is reapplied across recurrent updates, enabling fine-grained cross-modal interaction while limiting the task-specific attention parameter budget.
3. A training and evaluation design that separates controlled evidence degradation from class-imbalance assessment. Controlled-view supervision and direct-path distillation constrain the residual correction. Aggregate and class-level metrics, ablations, and degraded-input analyses then evaluate its contribution and bound the resulting claims.

The remainder of the paper reviews crisis-information classification and the relevant fusion mechanisms, formulates the general classification problem, describes LRIF-Net and its training procedure, reports the experimental evidence, and discusses the decision-support scope and limitations of the study.

2. Related Work

2.1. From Text-Based to Multimodal Crisis Information Classification

CrisisMMD provides a common benchmark for paired image-text disaster analysis [4]. Early multimodal work examined how textual and visual information contribute to humanitarian classification [5]. Text-focused systems, including CrisisBERT and Transformer-based multi-task learning, identify categories and requests from short messages [6,7], but do not exploit visual evidence in attached images.

Multimodal crisis classifiers used convolutional image encoders, including DenseNet and ResNet variants, together with text encoders [5,8]. Cross-modal bipolar attention, channel attention, and graph-based contextual reasoning increased fusion capacity for social-media analysis [9–11].

Recent crisis-specific studies use several routes to richer image-text evidence. CLIP-BCA-Gated applies bidirectional cross-attention and adaptive gating [12], and MKAN-Refine adds nonlinear refinement [13]. A caption-based model maps images into a text-compatible representation and examines missing modalities [14], while a recent CLIP study combines generated descriptions with guided and differential attention [15]. These methods strengthen crisis classification but do not provide the same combination of a parameter-tied latent residual and an explicitly retained direct classification path under the matched parameter-accounting protocol used here.

Several application-oriented studies show how automated analysis supports operational review. He and Hu combine disaster-post classification with an interactive geographic display [1]. Gao et al. extract damage information from text and images and analyze its spatial and temporal distribution during Typhoon Haikui [2]. These systems present classified posts or extracted damage information in forms that analysts can review. LRIF-Net examines the upstream classification component: local evidence fusion, compact adaptation, and the behavior of class predictions when evidence changes. Open-world disaster identification addresses the complementary problem of rejecting unknown categories [40].

2.2. Fine-Grained and Parameter-Efficient Image-Text Fusion

Fine-grained fusion has previously been implemented through Transformer attention, multiplicative pooling, co-attention, and multimodal Transformer variants [16–21].

Perceiver-style latent arrays compress high-dimensional inputs through a learned bottleneck and can reuse parameters across repeated computation, while attention-bottleneck models restrict cross-modal exchange through a small latent interface [22,23]. These established methods motivate, but are not themselves claimed as, new components of LRIF-Net.

Attention-guided bilinear fusion and pretrained vision-language models provide additional established mechanisms for linking textual cues with visual context [41–47]. Their relevance here is conceptual; evaluations on other tasks or protocols do not establish performance under the present humanitarian labels and split.

Parameter-efficient learning also has established precedents in low-rank matrices, adapters, feature scaling, and prompts [24–30]. LRIF-Net does not claim novelty for any one of these general principles. Instead, it uses parameter sharing to keep a recurrent evidence-correction path compact while leaving a complete direct classifier active in the final logit sum.

The specific contribution is the coordinated design of a decision-preserving recurrent residual framework for crisis classification. A complete retained classifier supplies the direct decision basis; the same compact interaction cell is reapplied to refine token-patch evidence; and controlled-evidence learning plus direct-path distillation constrain the zero-initialized residual output to act as an additive correction. The constituent operations are not claimed as new in isolation. Their coordination permits separate accounting of the direct and residual decisions and of the parameters added for refinement.

2.3. Incomplete Evidence and Decision Interpretation

Crisis posts may lack an image, contain little textual content, or pair text and imagery that refer to different situations. Dynamic-fusion and uncertainty-aware methods address low-quality or uncertain modalities [31–33]. Modality-recognition and modality-balancing studies examine reliability or dominance failures [34–36]. Modality dropout addresses unavailable inputs [37], training-difficulty analysis explains multimodal learning failures [38], and missing-modality learning handles severely missing modalities [39]. Crisis-specific caption-based work also evaluates incomplete modalities [14]. Here these conditions are controlled stress tests, not estimates of their prevalence in deployment.

Model interpretation is also relevant because emergency analysts need to understand which parts of a post influence a prediction. Attention weights alone need not be faithful explanations [48,49]. Perturbation methods, Grad-CAM, Integrated Gradients, and feature-removal benchmarks therefore motivate complementary tests of decision relevance [50–54].

A separate challenge is label-space shift. Leave-one-event-out evaluation changes the disaster event while retaining the five known categories, whereas open-world crisis classification explicitly rejects unknown categories [40]. The former is evaluated here as internal cross-event evidence; the latter remains outside the present model.

These considerations determine the evaluation: controlled input changes test reliance on surviving evidence, event-held-out folds test transfer among the observed disasters, and ranked feature interventions test which text and image elements affect the classification. Each test addresses a distinct part of the intended information-screening task.

2.4. Literature Synthesis and Research Gap

The reviewed work connects humanitarian classification to information retrieval and situational awareness, and it establishes latent bottlenecks, cross-attention, parameter sharing, residual updates, and controlled modality degradation as prior ideas. The remaining question is how these ideas can be coordinated around an already functioning crisis classifier so that local evidence is refined without discarding the direct decision or obscuring the additional adaptation cost.

LRIF-Net tests that question through a decision-preserving recurrent residual fusion framework. Its novelty claim is reserved for the coordinated architecture and evaluation—retained direct classification, compact parameter-tied recurrent residual interaction, and controlled-evidence learning for the evaluated crisis task—not for the constituent techniques in isolation. Table 1 therefore presents methodological context rather than a claim that each component is unprecedented.

Table 1. Methodological synthesis of crisis-social-media classification and compact multimodal fusion. The evidence boundary identifies why each family is contextual rather than a matched numerical comparator.

Study/Family	Crisis Task or Purpose	Fusion/Adaptation Mechanism	Evidence Boundary	Relation to LRIF-Net
CrisisMMD [4]	Benchmark humanitarian information categories	Paired image-text dataset	Dataset and task definition	Provides paired crisis data for the task
Abavisani et al. [8]	Crisis-event categorization	Instance-level multimodal attention	Different data and evaluation setup	Crisis-specific local alignment precedent
CrisisSpot [11]	Disaster-content classification	Graph reasoning with social context	Requires additional contextual graph	Alternative contextual evidence source
CLIP-BCA-Gated [12]	Humanitarian crisis classification	CLIP bidirectional cross-attention and gate	Different training/ reporting pipeline	Recent CLIP-based local fusion comparator
MKAN-Refine [13]	Fine-grained crisis classification	Cross-modal refinement and KAN layers	Different pipeline; no matched parameter protocol	Recent nonlinear crisis-fusion comparator
Yu et al. [40]	Known-class prediction and unknown rejection	Cross-attention plus open-world classifier	Different label-space objective	Shows the boundary of our closed-set task
Farah et al. [14]	Crisis-post categorization with incomplete modalities	Image captioning into a text-compatible space	Caption-based representation	Crisis-specific alternative to token-patch fusion
Munia et al. [15]	Three CrisisMMD classification tasks	CLIP, generated text, guided/differential attention	Different augmentation and protocol	Recent CLIP-based attention alternative
MCB/TFN/ LMF [18–20]	General multimodal prediction	Multiplicative or low-rank fusion	Not crisis-specific; pooled interactions	Motivates compact cross-modal interaction
MCAN/latent bottlenecks [21–23]	General vision-language fusion	Co-attention or learned latent interface	No retained crisis-classifier path	Motivates latent interaction; differs in path retention
PEFT methods [24–30]	Economical pretrained-model adaptation	Low-rank, adapter, scaling, or prompt updates	Not a crisis-specific fusion evaluation	Motivates task-specific parameter accounting
Robust and missing-modality studies [31–33,37,39]	Prediction with unavailable/ unreliable inputs	Dynamic fusion, uncertainty, dropout, rebalancing	Mostly general multimodal tasks	Motivates controlled evidence stress tests
LRIF-Net (ours)	Five-class crisis-information triage research prototype	Retained direct path plus two parameter-tied latent updates	Single closed-set benchmark; no operational validation	Tests the targeted performance/parameter trade-off

3. Materials and Methods

The experiments instantiate the general formulation below on the canonical CrisisMMD Task 2 split. After the canonical preparation [5], the evaluated label space contains five dataset-defined humanitarian information categories: affected individuals; infrastructure and utility damage; not humanitarian; other relevant information; and rescue, volunteering, or donation effort. The preparation merges injured/dead and missing/found

people into affected individuals, and vehicle damage into infrastructure and utility damage. These categories define the present closed-set experiment; they are not an exhaustive taxonomy of real-world crisis situations. Section 3.8 reports the split sizes and class counts.

3.1. Crisis Information Classification Problem

Consider a general closed-set multimodal classification task with an arbitrary number C of classes. Each observation is a paired social-media post comprising tokenized text and an image, together with one target label. The dataset is represented as follows:

$$\mathcal{D} = \{(xiT, xiV, yi)i = 1N\},\ yi \in \mathcal{Y} = \{1, \ldots, C\}. \quad (1)$$

Equation (1) defines the N paired observations and their labels. Its label space contains the category indices from 1 through C, where C is determined by the application rather than fixed by the formulation. For the CrisisMMD Task 2 experiment described above, C = 5. Given a paired post, the classifier must assign a score to every category and select the category with the largest score:

$$z_i = f_\theta(x_i^T, x_i^V) \in \mathbb{R}^C,\ \hat{y}_i = \arg\max_{c \in \{1,\ldots,C\}} z_{i,c}. \quad (2)$$

Equation (2) defines this input-to-decision mapping. The scoring function maps the text-image pair to a C-dimensional logit vector, and the predicted label is the index of its largest component. Thus, Equation (1) specifies the available observations and label space, whereas Equation (2) specifies the decision that must be learned from them.

LRIF-Net implements the scoring function through a complete direct classifier and an additive recurrent residual correction. Sections 3.2–3.7 describe how the two modalities are encoded, how the direct decision is retained, and how local cross-modal evidence is refined before the final logits are formed. The output dimension is set to C = 5 only for the present experiment. A different closed-set label space can use the same formulation with the corresponding value of C; dynamic labels or unknown-category rejection require separate objectives and evaluation [40].

3.2. LRIF-Net Architecture Overview

The architecture follows three requirements derived from the task formulation and literature synthesis. First, the classifier must retain both global post semantics and local word-patch evidence, because humanitarian categories can depend on whole-post context or localized cues. Second, fine-grained interaction should operate in a low-dimensional module with shared parameters, because unrestricted token-patch fusion increases the task-specific adaptation budget. Third, the learned interaction should refine rather than replace a complete direct classifier, so the direct logits and residual logits remain separately observable while forming the final prediction.

The subsections follow the sequence shown in Figure 2. Section 3.3 describes crisis-pair encoding, Section 3.4 the direct classification path, Section 3.5 recurrent evidence refinement, Section 3.6 the residual correction, and Section 3.7 the training-only objective.

The direct path supplies a complete prediction, while the latent path adds a residual logit vector. This direct-plus-residual coordination is the defining organization of the framework: recurrent token-patch refinement can modify the final decision without replacing the complete direct classifier, and both contributions remain available for separate analysis.

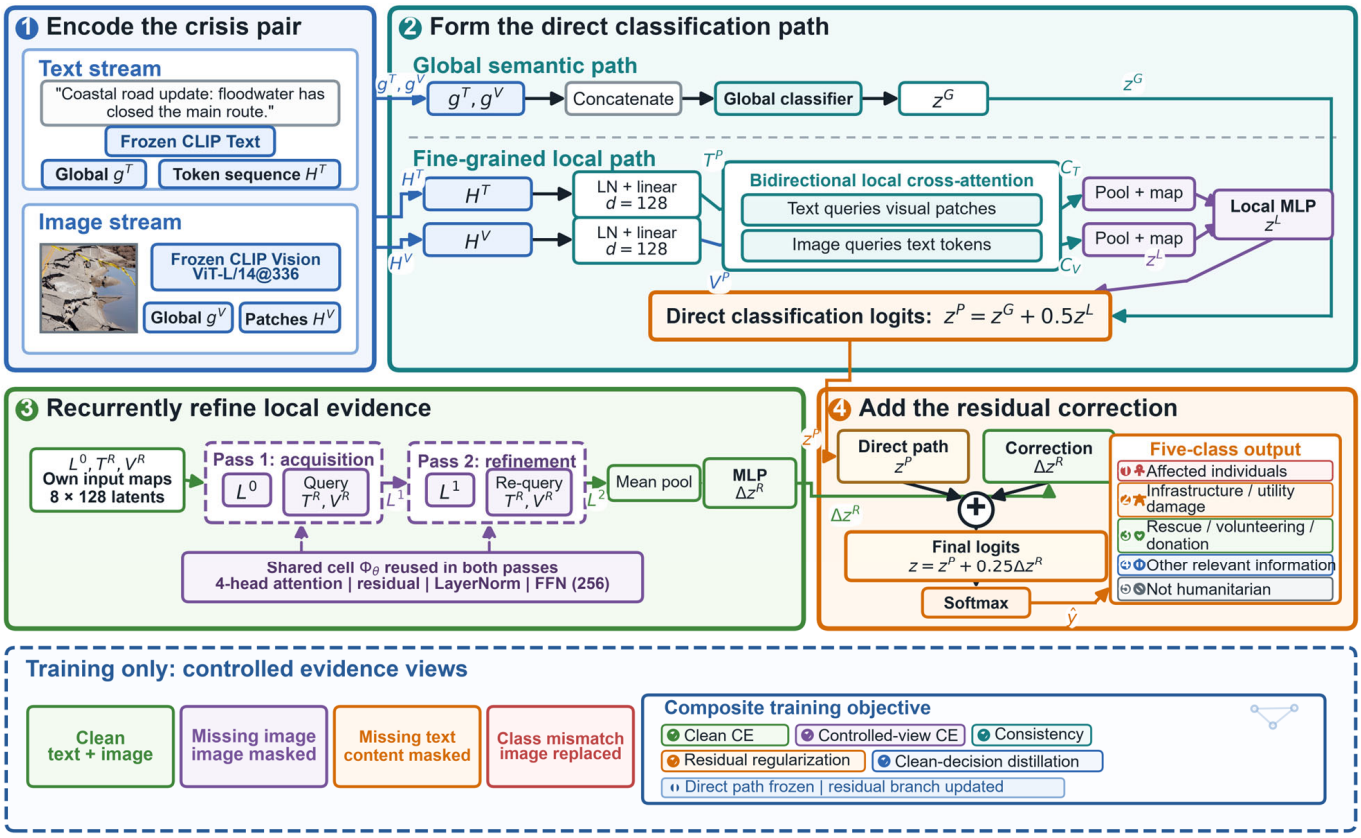


Figure 2. LRIF-Net architecture. Frozen CLIP encoders produce global embeddings and local token and patch sequences. The direct classification path forms $z^P = z^G + 0.5z^L$. Eight latent tokens perform two recurrent passes through the shared interaction cell Φθ to obtain Δz^R; $z = z^P + 0.25\Delta z^R$ yields the final prediction. The direct path remains frozen during recurrent training, and only the recurrent residual branch is updated. The lower band is training-only. The example image is a public-domain FEMA photograph by Robert Kaufmann.

3.3. Multiscale Crisis Evidence Encoding

In the 'Encode the crisis pair' block of Figure 2, frozen CLIP ViT-L/14@336 encoders [55] map text x^T and image x^V to global embeddings g^T and g^V and to local token and patch sequences H^T and H^V:

$$g^T, H^T = E_T(x^T), g^V, H^V = E_V(x^V). \quad (3)$$

Equation (3) defines the frozen global and local representations returned by the two encoders. The global embeddings summarize the complete post. In CLIP's vision Transformer, a learned [CLS] token summarizes the image sequence; non-[CLS] image patches are the remaining spatial patch embeddings. These patch embeddings retain localized evidence such as damaged structures, people, vehicles, or supplies and are used as keys and values in local cross-modal interaction. Ordinary text content tokens play the corresponding local linguistic role.

3.4. Global-Local Crisis Decision Formation

The 'Form the direct classification path' block in Figure 2 first L2-normalizes and concatenates the frozen CLIP global embeddings. The global head then applies LayerNorm, dropout, and a linear five-class projection to obtain the global logits in Equation (4):

$$z^G = f_G([g^T; g^V]). \quad (4)$$

Within the same Figure 2 direct-path block, the direct branch separately applies LayerNorm and a biased linear projection to text tokens and non-[CLS] image patches, as defined in Equation (5). Its projection functions map widths 768 and 1024, respectively, to d = 128; no learned modality embeddings are added in this branch:

$$T^P = \pi^P{}_T(H^T), V^P = \pi^P{}_V(H^V). \tag{5}$$

Still within the Figure 2 direct-path block, Equations (6) and (7) define two independently parameterized attention operators that produce image-conditioned text context and text-conditioned image context. They return the attention outputs without adding the input sequences:

$$C_T = \mathrm{MHA}_T(T^P, V^P, V^P), C_V = \mathrm{MHA}_V(V^P, T^P, T^P). \tag{6}$$

$$z^L = f_L([q_T; q_V]). \tag{7}$$

Text attention keys and text mean pooling exclude padding; valid boundary tokens remain. Each pooled context is mapped through LayerNorm, a linear projection from 128 to 768, and GELU, then L2-normalized to give q_T and q_V. The local head f_L is a LayerNorm–MLP with widths 1536→512→5, GELU, and dropout.

Here qT is the normalized output mapping of the masked mean of CT, and qV is the normalized output mapping of the mean of CV. The two output mappings have separate parameters. Equation (8) then combines the global and local heads to form the direct logits:

$$z^P = z^G + \alpha z^L, \alpha = 0.5. \tag{8}$$

This module is a complete multimodal crisis classifier. It remains active during inference and is frozen while the recurrent residual branch is trained, thereby providing the stable decision basis refined by LRIF-Net.

3.5. Shared Latent Evidence Interaction

The implementation separates shared-width alignment from the attention projections. Modality-specific LayerNorm–linear mappings first transform the CLIP text tokens and image patches to the shared width d = 128; these mappings do not create the head-specific query, key, and value representations. For a given attention call, its aligned query, key, and value inputs Q, K, and V are each projected exactly once for head j:

$$Q_j = QW_j^Q,\ K_j = KW_j^K,\ V_j = VW_j^V,\ j = 1, \ldots, H. \tag{9}$$

Here, the three W matrices are the learned query, key, and value projections for head j, while Qj, Kj, and Vj denote the resulting projected representations. The model uses H = 4 heads; therefore, each head has width dh = d/H = 32. Scaled dot-product attention is then computed only from Qj, Kj, and Vj:

$$h_j = \mathrm{softmax}(\frac{Q_j K_j^T}{\sqrt{d_h}})V_j. \tag{10}$$

After the four head outputs are concatenated, the attention module applies its standard output projection to return to the shared width. This post-concatenation operation is distinct from the query, key, and value projections in Equation (9); it does not project Qj, Kj, or Vj again. Accordingly, Equation (10) contains no weight matrix and no projected tensor is projected a second time. The text-, image-, and self-attention operators have separate parameters. Each sublayer has its own pre-normalization, whereas the complete recurrent

cell is reused across the two applications. In Equations (11)–(15), r indexes the recurrent application, t and v denote text- and image-conditioned states, j indexes attention heads, i indexes samples, and c indexes classes.

The recurrent inputs are $T^R = \pi^R{}_T(H^T) + e_T, V^R = \pi^R{}_V(H^V) + e_V$. Each πR is an independent LayerNorm–linear mapping; eT and eV are learned modality embeddings.

In Equations (11)–(13), N denotes LayerNorm, with separate parameters for the pre-normalizations in each sublayer.

$$A^r{}_t = L^{r-1} + \mathrm{MHA}_t(N(L^{r-1}), N(T^R), N(T^R)). \tag{11}$$

$$A^r{}_v = A^r{}_t + \mathrm{MHA}_v(N(A^r{}_t), N(V^R), N(V^R)). \tag{12}$$

$$S^r = A^r{}_v + \mathrm{MHA}_s(N(A^r{}_v), N(A^r{}_v), N(A^r{}_v)), L^r = S^r + \mathrm{FFN}(N(S^r)). \tag{13}$$

Equation (14) defines the first pass through the Figure 2 recurrent-refinement block, termed Initial Cross-Modal Evidence Acquisition. It forms an integrated latent description from the learned queries and the two encoded modalities:

$$L^1 = \Phi_\theta(L^0, T^R, V^R). \tag{14}$$

Equation (15) defines the second pass through the same Figure 2 recurrent-refinement block, termed Context-Conditioned Evidence Refinement. The updated state L^1 re-queries both modalities, allowing the second pass to refine evidence in the context collected by the first. Both applications use the same θ:

$$L^2 = \Phi_\theta(L^1, T^R, V^R), \theta_{\mathrm{acq}} = \theta_{\mathrm{ref}} = \theta. \tag{15}$$

Parameter tying deepens interaction without introducing a second attention-and-FFN parameter set. The ablation in Section 4.3 tests whether this contextual refinement contributes beyond the initial acquisition stage.

3.6. Direct-Path Residual Fusion

In the 'Add the residual correction' block of Figure 2, the refined latent state is pooled and mapped to the residual crisis logits in Equation (16):

$$\Delta z^R = f_R(\mathrm{MeanPool}(L^2)). \tag{16}$$

The residual head applies LayerNorm and a 128→128→5 MLP with GELU and dropout. Only its final linear layer is zero-initialized, giving an initial residual of zero. This initialization does not constrain the trained model to retain the direct argmax label. Equation (17) then adds the scaled residual correction to the direct logits to form the final logits:

$$z = z^P + \beta \Delta z^R, \beta = 0.25. \tag{17}$$

Path-level analysis compares the direct and final predictions and decomposes class-contrast changes into direct and scaled residual contributions. This quantifies how the retained classification path and learned correction contribute to the final decision.

3.7. Controlled-Evidence Learning Objective

The lower training-only band in Figure 2 combines the clean pair with one controlled view: a zero image tensor in CLIP-normalized pixel space, text with ordinary content masked while retaining boundary tokens, or an image drawn from a different class in the same split. Every view retains the original text-side target label. Equation (18) defines these

controlled transformations; image-mismatch partners are generated deterministically using the run seed for training and seed + 10,000 for validation:

$$c_k(x^T, x^V) \in \{(x^T, 0), (Boundary(x^T), x^V), (x^T, x^V{}_j) : y_j \neq y\}. \quad (18)$$

The objective in Figure 2 training-only band combines clean classification, controlled-view classification, clean-to-view consistency, residual regularization, and distillation from the direct decision path [56]. Equation (19) gives the composite objective, and Equation (20) defines its consistency and distillation terms:

$$L = L^{clean}{}_{CE} + 0.5L^{corrupt}{}_{CE} + 0.1L_{cons} + 0.01R_{res} + 2.0L_{KD}. \quad (19)$$

$$L_{\text{cons}} = D_{\text{KL}}\left(\text{sg}(p_{\text{clean}}) \| p_{\text{corrupt}}\right), L_{\text{KD}} = D_{\text{KL}}(p_P \| p_{\text{clean}}). \quad (20)$$

The loss coefficients were empirically and heuristically prespecified before the formal experiments, based on the intended roles and relative numerical scales of the five loss terms, and were then kept fixed. They were not obtained through a grid search, random search, candidate-value sweep, or other systematic validation experiment. The coefficients are 1.0 for clean cross-entropy, 0.5 for corrupted-view cross-entropy, 0.1 for clean-to-view consistency, 0.01 for residual regularization, and 2.0 for direct-path distillation. Clean cross-entropy carries the main task signal, while the other terms provide auxiliary controlled-view supervision and soft constraints. The clean distribution is detached in the consistency term, the direct teacher is frozen, both KL terms use a temperature of 1, and residual regularization averages the squared unscaled residual logits equally over clean and corrupted views, samples, and classes. Validation Accuracy was used only to select the checkpoint epoch; neither the validation results nor the test set was used to select or adjust these coefficients.

In Equation (20), sg denotes stop-gradient. Softmax probabilities are computed from the corresponding clean, corrupted, or frozen direct logits. The corruption type cycles once per optimizer step; four accumulated microbatches use the same type, and the cycle restarts each epoch.

3.8. Dataset, Evaluation Metrics, and Statistical Analysis

The canonical CrisisMMD Task 2 subset contains 7314 label-consistent pairs: 5119 training, 1097 validation, and 1098 test examples (Table 2). Following the five-class preparation [5], injured/dead and missing/found people are merged into affected individuals, and vehicle damage is merged into infrastructure and utility damage; pairs with inconsistent modality labels are excluded. Checkpoint selection and member-level evaluation are detailed in Supplementary Section S10.

Table 2. Canonical CrisisMMD Task 2 split used in all formal comparisons.

Class	Train	Validation	Test	Total
Affected individuals	81	17	17	115
Infrastructure and utility damage	497	106	107	710
Not humanitarian	2651	568	568	3787
Other relevant information	1164	250	250	1664
Rescue, volunteering, and donation	726	156	156	1038
Total	5119	1097	1098	7314

For C classes, let TPc, FPc, TNc, and FNc denote the class-wise true-positive, false-positive, true-negative, and false-negative counts; nc is the number of samples in class c and N is the total sample count. The evaluation uses a prespecified set of complementary metrics rather than a single primary score. Accuracy, Macro F1, Weighted F1, per-class

precision, recall, and F1, and the confusion matrix are reported for the main comparisons, with Balanced Accuracy as an additional class-balanced diagnostic. Weighted F1 summarizes performance under the observed class frequencies and is not used alone to claim predictive superiority:

$$Accuracy = \frac{\sum_{c=1}^{C} TP_c}{N}. \tag{21}$$

$$Precision_c = \frac{TP_c}{TP_c + FP_c}, Recall_c = \frac{TP_c}{TP_c + FN_c}. \tag{22}$$

$$F1_c = \frac{2Precision_c Recall_c}{Precision_c + Recall_c}. \tag{23}$$

$$MacroF1 = \frac{1}{C}\sum_{c=1}^{C} F1_c. \tag{24}$$

$$WeightedF1 = \sum_{c=1}^{C} \frac{n_c}{N} F1_c. \tag{25}$$

$$\text{Balanced Accuracy} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}. \tag{26}$$

Equations (21)–(26) define Accuracy, class-wise Precision and Recall, class-wise F1, Macro F1, Weighted F1, and Balanced Accuracy, respectively. Accuracy is the overall proportion of correct predictions. Macro F1 gives every class equal weight, whereas Weighted F1 weights each class F1 by its sample support. Balanced Accuracy is the arithmetic mean of the class recalls. Per-class F1 and the confusion matrix identify category-specific performance and error directions. Because the five classes are imbalanced, no one metric is treated as sufficient evidence of predictive superiority. Clean Validation Accuracy is used only to select checkpoints; after the model and checkpoints are fixed, the test results are interpreted jointly across the aggregate and class-level metrics.

The canonical test analysis uses all 1098 test posts and the fixed three-seed probability ensembles for seeds 3141, 1729, and 2718; each ensemble is the arithmetic mean of the three softmax probability vectors with weights of 1/3. For each fixed-system comparison, 10,000 paired bootstrap resamples are drawn within the true classes. The post is the resampling unit, and the same sampled post indices are used for both systems. The 2.5th and 97.5th percentiles of the paired metric differences form the 95% percentile interval. The bootstrap random seed is 20260725. Exact two-sided McNemar tests use the discordant correctness counts from the same paired test predictions and the conditional probability 0.5 under equal marginal accuracy. These post hoc comparisons are exploratory and are not adjusted for multiple comparisons; they condition on the observed class counts and fitted models and do not estimate variation from new training runs. They assume independent posts, with within-class exchangeability for the bootstrap; event and repeated-content dependence were not adjusted for. Sample IDs and labels were aligned before analysis. Supplementary Sections S10 and S12 provide the checkpoint, evaluation-sequence, and saved-comparison details.

3.9. Implementation and Comparators

All formal models use the frozen openai/clip-vit-large-patch14-336 backbone and seeds 3141, 1729, and 2718. The architecture uses eight latent tokens as a compact bottleneck, a width of 128 to match the direct local-interaction space, and four attention heads. The shared cell is applied twice, first for evidence acquisition and then for contextual refinement, without introducing a second parameter set. These settings were fixed design choices. A post hoc validation-only sensitivity analysis recombines saved logits without retraining: alpha takes 0.25, 0.50, and 0.75 at beta = 0.25, and beta takes 0.125, 0.25, and

0.375 at alpha = 0.50. The equal-probability three-seed analysis describes local coefficient sensitivity and does not select or alter the locked model.

For each seed, the first epoch attaining the highest clean Validation Accuracy selects the checkpoint. The ensemble averages the three softmax probability vectors equally. Controlled baselines comprise frozen global CLIP, the direct 128-dimensional four-head global-local classifier, and its 512-dimensional eight-head cross-attention counterpart. The LRIF-Net method, checkpoints, coefficients, and ensemble weights were fixed before its test evaluation and remained unchanged afterward. Supplementary Section S10 records checkpoint selection and the evaluation sequence.

A single LRIF-Net member runs the frozen direct path and two applications of the latent cell, then adds their logits. Test predictions average the probabilities of three members. The parameter and timing analyses distinguish one-member measurements from complete-ensemble measurements. Final-stage training uses AdamW (learning rate and weight decay both 10^{-4}), batch size 1 with four-step gradient accumulation, at most eight epochs, and an early-stopping patience of four epochs. No class weights are used. The implementation uses PyTorch 2.5.1+cu121 and Transformers 4.45.2. Supplementary Sections S8–S10 specify architecture, preprocessing, training, parameter boundaries, and member-level results; Section S15 details ensemble execution.

4. Results

The Results first report classification performance and class-level errors (Sections 4.1 and 4.2), then examine incomplete evidence and event transfer (Sections 4.3 and 4.4), parameter and execution costs (Section 4.5), and features that influence individual decisions (Section 4.6).

4.1. CrisisMMD Classification Performance

Table 3 situates LRIF-Net among representative CrisisMMD Task 2 methods. The historical rows reproduce the consolidated scores reported by Li et al. [12] and provide literature context because their splits and training protocols differ, while Section 4.2 supplies the matched-protocol comparison.

Table 3. Literature context for CrisisMMD Task 2. Values are percentages from the respective published protocols; historical Precision, Recall, and F1 retain their source definitions. LRIF-Net Precision/Recall are micro averages and F1 is Weighted F1. Dashes denote unavailable values. Matched comparisons are reported separately in Table 4.

Method	Backbone	Protocol Status	Acc.	Prec.	Recall	F1
VGG-16 + CNN [12]	CNN	Reported; differing setup	78.40	78.50	78.00	78.30
DenseNet + BERT [12]	DenseNet/BERT	Reported; differing setup	82.72	82.50	82.72	82.46
ResNet-152 + BERT [12]	ResNet/BERT	Reported; differing setup	84.10	84.13	84.10	83.92
CrisisBERT [12]	BERT	Reported; differing setup	85.60	85.74	85.60	85.58
ViT + BERT [12]	ViT/BERT	Reported; differing setup	86.53	86.62	86.53	86.21
FBP [12]	Multimodal	Reported; differing setup	—	88.50	88.10	88.10
CBAN-Dot [12]	Multimodal	Reported; differing setup	88.38	87.95	87.80	88.40
DMCC [12]	Multimodal	Reported; differing setup	88.00	87.95	87.80	87.72
CrisisSpot [12]	ViT/BERT/GNN	Reported; extra context	88.65	88.74	88.65	88.32
CLIP-BCA-Gated [12]	CLIP	Reported; differing setup	91.77	—	—	91.74
MKAN-Refine [13]	CLIP	Reported; differing setup	92.08	—	—	92.07
LRIF-Net (ours)	CLIP ViT-L/14@336	Canonical 7314 split	91.35	91.35	91.35	91.32

For context, CLIP-BCA-Gated reports 91.77% Accuracy and 91.74% F1 [12], and MKAN-Refine reports 92.08% Accuracy and 92.07% Weighted F1 [13]. These published

values use different data construction, augmentation, training, and model-selection procedures. Table 4 evaluates the present design under matched data, backbone, and checkpoint-selection conditions.

4.2. Matched Crisis Classification Results

Table 4 reports the matched canonical comparison under the shared evaluation protocol.

Table 4. Three-seed ensembles evaluated under the identical canonical protocol. Bold values indicate the highest numerical result within this controlled table.

System	Accuracy	Macro F1	Weighted F1
Frozen global CLIP	0.9062	0.8782	0.9058
Preserved global-local model	0.9126	**0.8952**	0.9124
Standard cross-attention	0.9089	0.8912	0.9088
LRIF-Net	**0.9135**	0.8888	**0.9132**

LRIF-Net records the highest numerical Accuracy (91.35%) and Weighted F1 (91.32%) in Table 4, exceeding standard cross-attention by 0.46 and 0.44 percentage points. However, its Macro F1 is 88.88%, compared with 89.52% for the retained direct global-local model, and its Balanced Accuracy is 86.96% versus 87.27%. Supplementary Tables S1 and S2 report per-class metrics and the confusion matrix. These differences across metrics motivate the class-level and paired analyses below rather than a claim of uniform predictive improvement.

For the affected-individuals class, both systems correctly classify 12 of 17 posts. One additional false positive lowers LRIF-Net F1 from 82.76% for the direct model to 80.00%. Of the five posts missed by both systems, three are assigned to not humanitarian, one to infrastructure and utility damage, and one to rescue, volunteering, or donation. Across all classes, LRIF-Net gains four correct predictions for not humanitarian and loses two for infrastructure and utility damage and one for other relevant information, yielding 1003 versus 1002 correct posts. Thus, the aggregate gain does not extend to improved minority-class recognition.

Direct and final labels agree on 1079 of 1098 posts (98.27%). The 19 changes comprise nine direct-only and 10 LRIF-Net-only correct predictions (exact McNemar $p = 1.0$). Paired differences are +0.09 percentage points in Accuracy (95% CI −0.73 to +0.91), −0.63 in Macro F1 (−2.11 to +0.40), and +0.08 in Weighted F1 (−0.72 to +0.88). Accordingly, LRIF-Net provides competitive performance under the matched evaluation protocol, but these differences do not establish predictive superiority; Section 4.5 separately quantifies the adaptation-parameter benefit.

The saved-prediction comparison with standard cross-attention gives Accuracy +0.46 points [−0.55, +1.46], Macro F1 −0.24 [−1.79, +1.00], and Weighted F1 +0.44 [−0.54, +1.42]; exact McNemar $p = 0.4583$ (17 LRIF-Net-only and 12 comparator-only correct). The global-CLIP intervals also cross zero. Supplementary Table S13 reports all three comparisons under the statistical protocol in Section 3.8.

Post Hoc Validation Coefficient Sensitivity

With the frozen models, the nominal alpha = 0.50 and beta = 0.25 ensemble obtains 91.25% Accuracy, 87.74% Macro F1 and 91.20% Weighted F1 on validation data. Across the five local settings in Supplementary Table S8, the maximum absolute changes are 0.73, 0.60 and 0.73 percentage points, respectively. Lowering alpha to 0.25 raises Macro F1 by 0.31 points while reducing Accuracy by 0.36 points. Thus the settings trade off metrics. This post hoc recombination of fixed logits does not evaluate retraining or establish optimal coefficients; the locked test model is unchanged.

4.3. Controlled Evidence-Degradation and Component Analysis

The controlled evidence-degradation analysis first compares LRIF-Net with its frozen direct global-local source under three predefined synthetic conditions: missing image, missing text, and different-class image mismatch. Under the registered corruption-aware objective, paired Validation Accuracy exceeds that source by 20.30 points, 6.87 points, and 18.87 points, respectively; the equal-condition mean is 15.34 points. These values characterize the evaluated protocols and do not establish robustness to all naturally occurring missing, noisy, or misleading crisis data.

These gains compare the complete system with its frozen source and therefore combine architecture and training effects. Table 5 examines component removal, and Table 6 compares systems trained with the same corruption schedule and loss coefficients, subject to the teacher and regularization differences stated below.

Table 5. Three-seed validation ablation. Controlled-degradation gain is the equal-condition Accuracy gain over the seed-matched direct global-local classifier, in percentage points.

Variant	Clean Acc. (%)	Clean Drop	Controlled-Degradation Gain	Controlled-Degradation Drop
Full LRIF-Net	90.79 ± 0.33	—	15.34	—
Without clean-decision distillation	90.46 ± 0.05	0.33	12.64	2.70
Without corruption objective	91.01 ± 0.19	−0.21	0.00	15.34
Initial evidence acquisition only	90.70 ± 0.40	0.09	14.29	1.05
Without text cross-attention	90.64 ± 0.19	0.15	6.74	8.61
Without image cross-attention	91.10 ± 0.23	−0.30	5.54	9.80
Without latent self-attention	90.70 ± 0.27	0.09	12.51	2.84
Without feed-forward network	90.61 ± 0.40	0.18	12.47	2.88

Table 6. Three-seed validation controls with the same corruption cycle, loss coefficients, optimizer, training budget, and checkpoint rule. Accuracy is mean ± sample standard deviation. The teacher and the regularized logit branch are architecture-specific, so this is not a fully isolated architectural comparison. Parameters are those updated in the final control stage.

System	Control-Stage Trainable Parameters	Clean	Missing Image	Missing Text	Image Mismatch	Mean Degraded
Direct global-local + objective	1,356,549	89.97 ± 0.33	83.62 ± 0.97	86.36 ± 0.19	62.59 ± 4.39	77.52 ± 1.68
Standard cross-attention + objective	4,605,957	90.61 ± 0.33	81.01 ± 1.74	86.17 ± 0.19	52.66 ± 6.50	73.28 ± 2.11
LRIF-Net	517,509	90.79 ± 0.33	79.22 ± 1.88	84.08 ± 0.43	50.47 ± 9.52	71.25 ± 3.82

Table 5 evaluates the contribution of each component to clean classification and adaptation under incomplete evidence while retaining the remaining configuration.

Removing corruption-aware training reduces the mean source-relative controlled-degradation gain from 15.34 to 0.00 points, while clean Accuracy rises from 90.79% to 91.01%. Removing image cross-attention raises clean Accuracy to 91.10% but lowers the corresponding gain to 5.54 points. Relative to one application, the second shared application increases mean clean Accuracy by 0.09 points and the controlled-degradation gain by 1.05 points without adding a second cell parameter set. These descriptive ablations identify the learning objective as the main driver under the three specified conditions and show the clean/degraded-input trade-off of the retained components.

LRIF-Net combines the highest mean clean Validation Accuracy in Table 6 with the fewest updated parameters. The direct global-local control has the highest absolute Accuracy in all three controlled degradation conditions: mismatch Accuracy is 62.59 ± 4.39%, compared with 50.47 ± 9.52% for LRIF-Net. The compact residual therefore offers an

adaptation-cost trade-off rather than the strongest absolute degraded-input performance. Each architecture uses its own source teacher; L2 regularizes the reopened local logits in the controls and the recurrent logits in LRIF-Net, as detailed in Supplementary Section S3.

4.4. Within-Benchmark Leave-One-Event-Out Evaluation

Figure 3 evaluates each of seven held-out events over three seeds, using only the 6216 canonical training-plus-validation examples. For each fold, the global parent, direct local module, and recurrent branch are trained anew without the held-out event. The remaining original training rows are used for fitting and the remaining original validation rows for checkpoint selection; all train-plus-validation rows from the held-out event are reserved for evaluation. Canonical test examples are not used. Supplementary Sections S4 and S11 provide the split construction and absolute results.

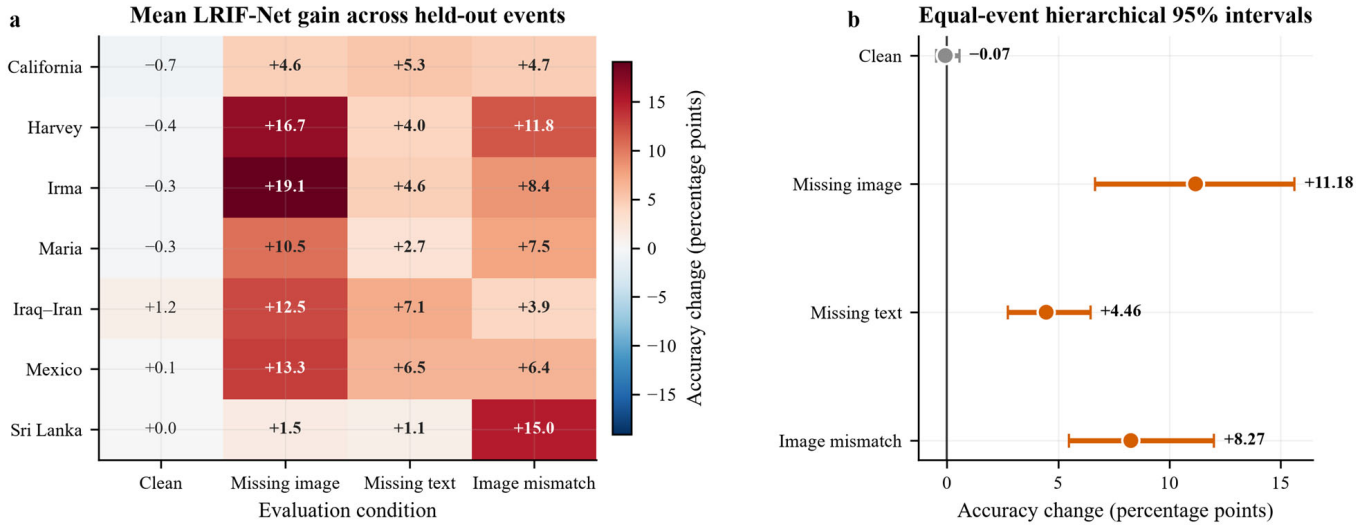


Figure 3. Leave-one-event-out transfer within the seven-event CrisisMMD benchmark and across three seeds. (**a**) Event-wise Accuracy changes under clean and controlled evidence conditions; (**b**) equal-event mean changes with hierarchical 95% confidence intervals.

Figure 3 shows that equal-event mean changes on complete pairs remain near zero (−0.07 points in Accuracy, +0.19 points in Macro F1, and −0.05 points in Weighted F1), while missing-image, missing-text, and mismatch Accuracy gains reach 11.18, 4.46, and 8.27 points. Their equal-weight mean is 7.97 points (95% CI: 6.18–9.72; exact event-level sign-flip $p = 0.015625$).

All seven event-level mean corruption gains are positive, showing consistent source-relative improvement across the evaluated disasters. These results characterize transfer within CrisisMMD; the seven-event design and overlapping training folds limit this conclusion to within-benchmark transfer.

4.5. Parameter Efficiency and Inference Cost

Figure 4 compares parameters updated in the final training stage, using the actual checkpoint-matched model structures. Supplementary Table S10 separately reports all non-CLIP parameters and the complete inference model.

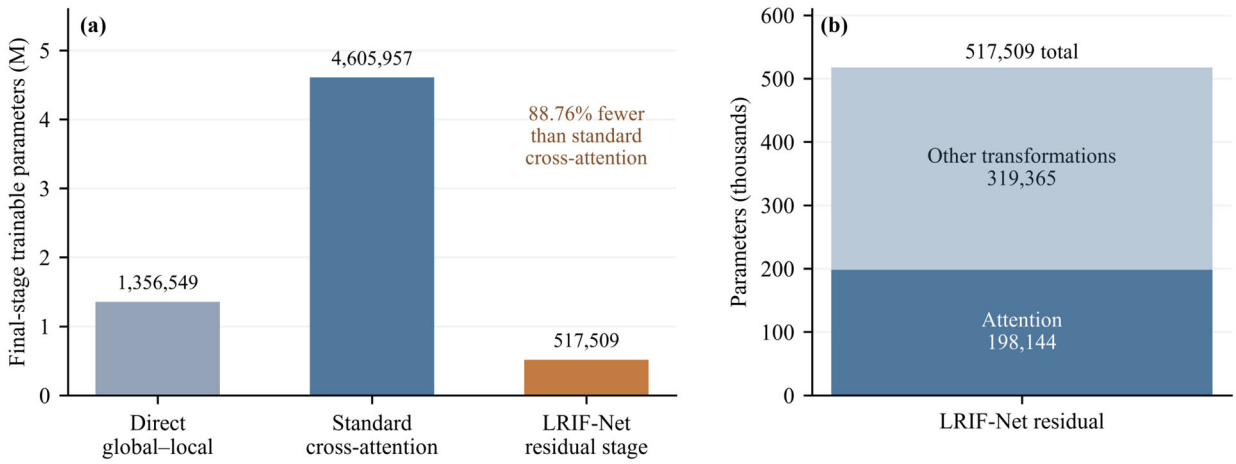


Figure 4. Final-stage parameter accounting for one model member. (**a**) The direct global-local control, its 512-dimensional cross-attention counterpart, and the LRIF-Net residual stage. (**b**) LRIF-Net residual parameters: attention operators, 198,144; other transformations, 319,365. Frozen components remain necessary at inference and are excluded only from these update counts.

Table 7 reports single-member throughput for the direct model, standard cross-attention, and LRIF-Net. The six cyclic/reverse-order trials used one Tesla V100-SXM2-32 GB GPU (NVIDIA Corporation, Santa Clara, CA, USA), mixed precision, 20 warm-up and 100 measured forwards per trial. The timed region excludes model loading and data preprocessing.

Table 7. Median fixed-V100 throughput. Values are samples per second.

Model	Batch 1	Batch 8	Batch 16	Scope
Preserved global-local model	29.00	94.50	108.54	Same V100 protocol
Standard cross-attention	29.22	93.63	107.64	Same V100 protocol
LRIF-Net	26.11	104.48	114.47	Same V100 protocol
LRIF vs. direct	−11.05%	+10.56%	+5.46%	Descriptive device-specific result

LRIF-Net updates 517,509 parameters over its frozen direct parent, compared with 4,605,957 for standard cross-attention (88.76% fewer). It retains 1,884,815 non-CLIP parameters versus 4,616,714 for standard attention and 1,367,306 for the direct model. Supplementary Section S9 gives the component-level accounting.

The residual branch comprises 198,144 attention and 319,365 other parameters. Including CLIP, LRIF-Net has 429,829,008 parameters versus 432,560,907 for standard attention: only 0.63% fewer. Final-stage update counts exclude prior parent training.

On this V100, LRIF-Net is slower than the direct model at batch 1 and faster at batches 8 and 16. Throughput therefore depends on batch size and hardware; the parameter reduction specifically concerns task adaptation.

To quantify the cost of ensemble inference, Table 8 measures the complete three-member direct and LRIF-Net systems on an RTX 3060 Laptop GPU. Four balanced trials use 16 warm-up and 64 measured requests per condition, cycling through the first 16 validation posts at batch sizes 1 and 8. Frozen CLIP weights share storage after exact tensor-identity checks, while each request executes three complete forwards and averages their softmax probabilities. Prepared-forward timing and decoded-input-to-host-probability timing are measured separately; the latter includes preprocessing and GPU transfers from already decoded images and raw text (Supplementary Section S15).

Table 8. Three-member ensemble execution on an RTX 3060 Laptop GPU. Latency is the median of four trial means [minimum, maximum]; throughput is the median of trial throughputs. Memory is the maximum PyTorch 2.5.1+cu121 allocated peak (MiB = 2^{20} bytes). F: prepared forward; D: decoded input to host probabilities. All cases use shared frozen CLIP storage and three full forwards.

Model	Batch	Scope	Latency (ms/Batch) Median [Range]	Throughput (Samples/s)	Allocated (MiB)
Direct	1	F	234.82 [142.16, 258.49]	4.30	1719.84
Direct	1	D	227.43 [153.26, 250.05]	4.40	1689.72
Direct	8	F	1325.46 [711.83, 1473.35]	6.04	1865.38
Direct	8	D	1256.17 [881.19, 1404.71]	6.42	1855.38
LRIF-Net	1	F	205.58 [158.00, 234.87]	4.86	1721.47
LRIF-Net	1	D	211.41 [168.60, 212.53]	4.73	1691.45
LRIF-Net	8	F	1108.11 [988.65, 1223.76]	7.29	1867.49
LRIF-Net	8	D	1120.93 [1024.29, 1258.03]	7.19	1857.15

For LRIF-Net, decoded-input execution takes a median 211.41 ms per single-post request and 1120.93 ms per eight-post request, with corresponding allocated peaks of 1691.45 and 1857.15 MiB. This provides a measured memory and latency budget for the reported ensemble on the evaluated device. The ranges show substantial variation between trials, so these measurements describe execution cost rather than a stable speed advantage.

4.6. Decision-Relevant Image and Text Evidence

4.6.1. Path-Level Image Evidence Visualization

Figure 5 shows five fixed validation examples of patch-occlusion evidence for the direct, recurrent, and final paths. These cases are qualitative; the separate deletion/insertion analysis uses larger class-stratified samples.

For each 14 × 14 input patch, the occlusion analysis measures the change in the true-class contrast: the target logit minus the mean of the other four logits. The residual score uses the scaled correction, $\beta \Delta z_R$. Consequently, direct and residual score changes add to the final score change before display processing. The 24 × 24 maps are bilinearly enlarged and Gaussian-smoothed; positive direct/final maps and signed residual maps are scaled separately by their 98th-percentile positive or absolute magnitude, respectively, and clipped. Independent display scaling prevents an absolute comparison of colors across paths.

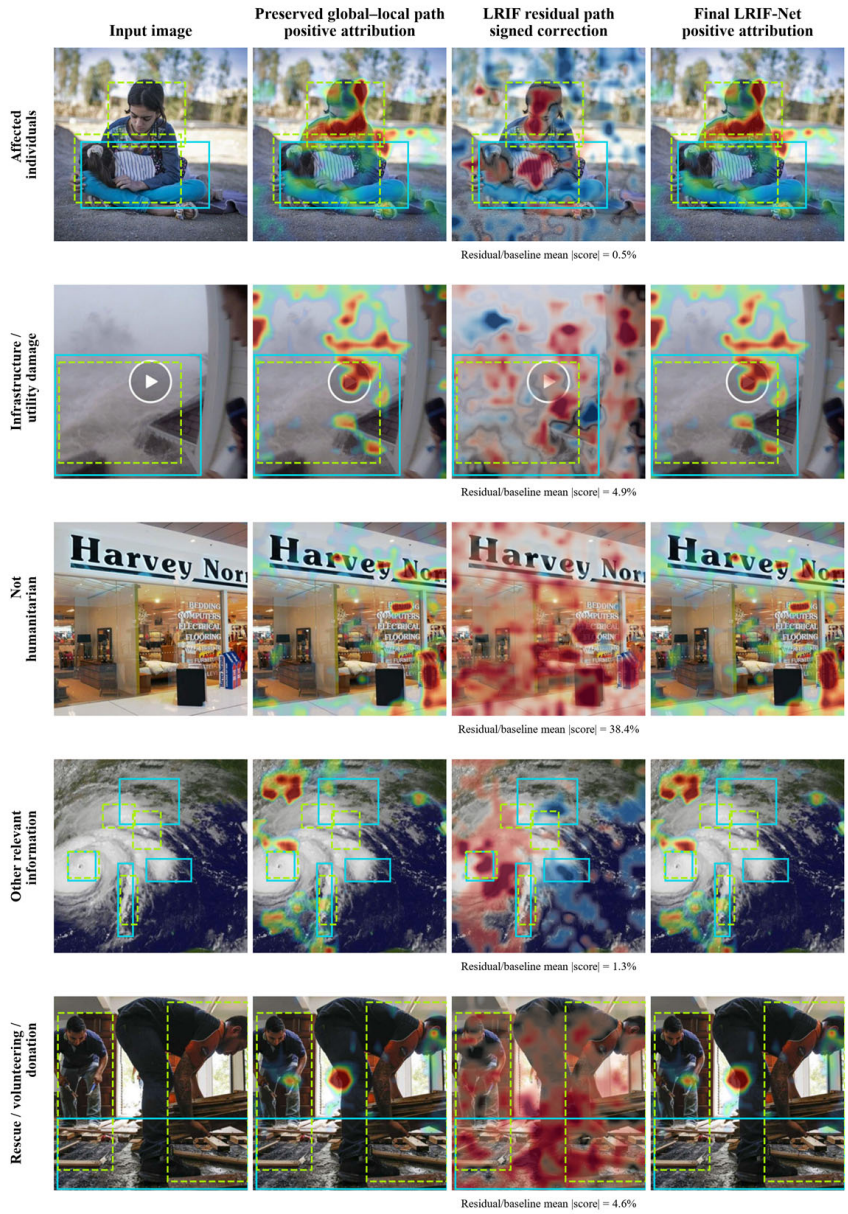


Figure 5. Author-generated patch-occlusion visualizations for five fixed CrisisMMD validation examples [4]. Columns: input, positive direct contribution, signed recurrent correction (red positive, blue negative), and positive final contribution. Cyan solid and lime dashed rectangles mark the two annotators' reference regions. Each map is normalized separately. Residual/direct mean absolute score ratios are 0.5%, 4.9%, 38.4%, 1.3%, and 4.6% from top to bottom, which helps explain the similarity of several direct/final panels.

4.6.2. Deletion–Insertion Decision-Relevance Analysis

Figure 6 evaluates text interventions across all 1097 canonical validation posts with the frozen seed-3141 models, including incorrect predictions. No post was excluded for empty content. Top-ranked, bottom-ranked, and five random token orders are evaluated at 10%, 30%, and 50% deletion or insertion. The primary target is each model's clean predicted class; true-class effects and correctness/class strata are reported in Supplementary Section S13. Nested token sets and shared random permutations provide matched controls across fractions and models.

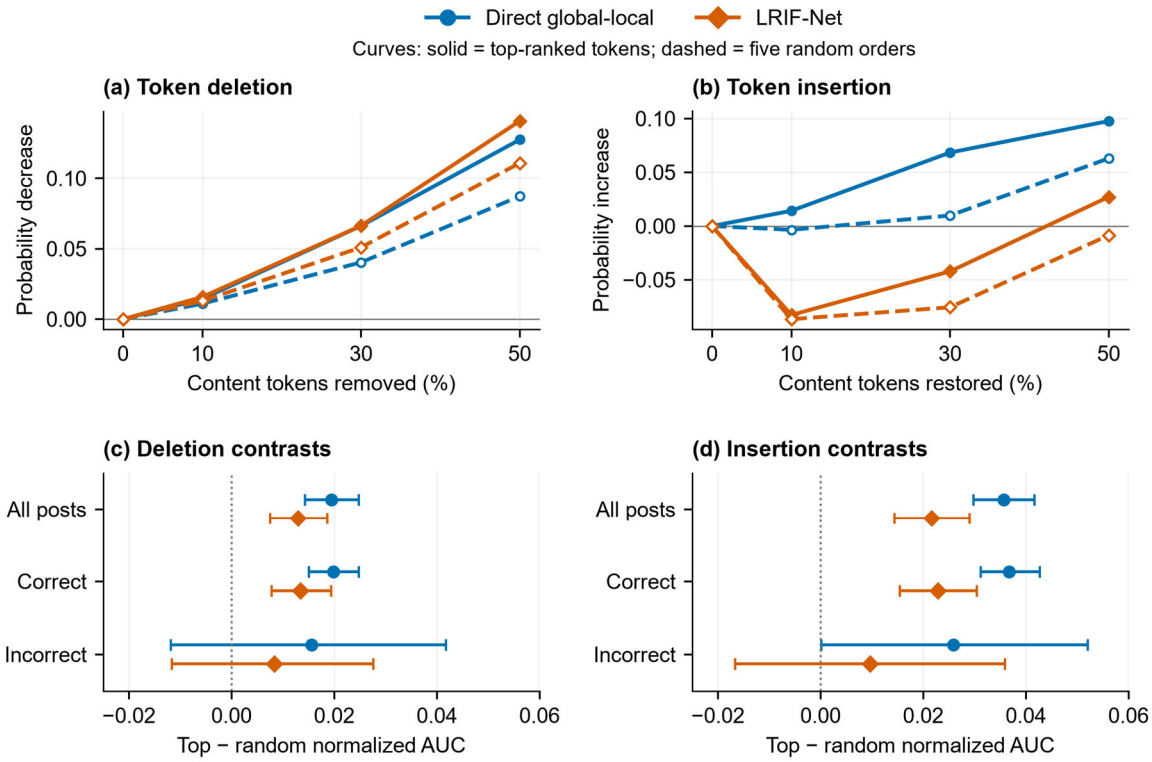


Figure 6. Text interventions on all 1097 validation posts. (**a**,**b**) Mean changes in clean predicted-class probability under token deletion/insertion; solid lines show top-ranked tokens and dashed lines average five random orders. (**c**,**d**) Mean top-minus-random normalized AUC with 95% intervals from 10,000 paired, class-stratified bootstrap resamples. Correct/incorrect subsets contain 993/104 posts for the direct model and 995/102 for LRIF-Net, defined by this diagnostic pass. Intervals compare token orders within each model. True-class and per-class results appear in Tables S14 and S15.

Across the full validation population, LRIF-Net has positive top-minus-random contrasts for deletion (+0.0130, 95% CI [0.0074, 0.0186]) and insertion (+0.0217 [0.0143, 0.0290]). These results extend the evidence for decision-relevant text rankings beyond the original correct-prediction subsets. The direct model also has positive contrasts, with larger point estimates (Table S14). Within LRIF-Net's 102 incorrectly classified posts, both intervals include zero. Thus the aggregate effect does not establish reliable explanations of individual errors.

The original class-capped text study remains separate in Supplementary Figure S1 and Section S5; its 170-post selection and randomization differ from the full-population extension. The image study also retains its original 170 correctly classified posts per model. LRIF-Net top-minus-random patch-ordering AUC is +0.0531 [0.0374, 0.0700] for deletion and +0.0695 [0.0524, 0.0875] for insertion. Together, the image and text interventions identify features affecting class scores under the specified operators.

5. Discussion

5.1. Predictive Performance and Adaptation Cost

The main finding is that a crisis classifier can be adapted through a small recurrent correction while retaining competitive performance under the matched evaluation protocol. LRIF-Net has the highest numerical Accuracy and Weighted F1 in Table 4, but its Macro F1 and Balanced Accuracy are lower than those of the retained direct classifier, and paired intervals include zero. The residual stage updates 517,509 parameters, 88.76% fewer than matched standard cross-attention; including the retained parent gives a 59.17% reduction in

all non-CLIP parameters. This is an attention-component and adaptation-parameter benefit with measured inference costs, not evidence that the full CLIP system is small or faster.

The class-level results explain what aggregate performance means for humanitarian screening. The not-humanitarian majority contributes strongly to Accuracy and Weighted F1, whereas Macro F1 exposes performance across all five categories. The larger categories also attain higher F1, so weighting by support raises Weighted F1 above the unweighted mean. Both the direct and final systems correctly classify 12 of the 17 posts in the affected-individuals class. The residual branch therefore changes some decisions without improving recall for this rare category. An analyst-facing evaluation should retain such category-specific checks alongside the aggregate metrics.

The ablations connect gains under the three controlled evidence-degradation conditions primarily to the learning objective. Recurrent latent interaction provides a compact way to apply that objective while retaining a complete direct classifier. The matched-objective controls identify the operating trade-off: LRIF-Net favors clean performance and fewer updates, while the larger direct control attains higher absolute Accuracy under each synthetic degraded-input condition. Consequently, these experiments bound behavior under the registered protocols rather than establishing robustness to arbitrary real-world crisis data.

5.2. Scope of the Contribution

Latent bottlenecks, cross-attention, parameter sharing, recurrent refinement, residual connections, and controlled modality degradation have all been used previously [12–39]. LRIF-Net therefore claims no isolated component as novel. Its system-level contribution is the coordinated decision-preserving recurrent residual design. A complete direct classifier remains active as the decision basis, and a compact parameter-tied cell is reapplied to refine token-patch evidence. Controlled-evidence supervision and direct-path distillation constrain a zero-initialized additive correction instead of a replacement prediction. Together, these choices address fine-grained interaction, decision preservation, and adaptation cost for the evaluated crisis-classification task.

5.3. Interpretation of the Residual Recurrent Design

The direct/residual decomposition links the model structure to the observed predictions. Most final predictions agree with the direct classifier, while the changed cases reveal where the learned correction alters the decision. Patch-occlusion maps further separate the contribution of each path in the same class-contrast units before display normalization.

Reusing the same cell lets the second application revisit image-text evidence with context acquired in the first. The modest ablation differences support studying recurrence as refinement within a fixed parameter budget, rather than adding another wide interaction layer. The present two-application setting is a specific design choice, not an optimized depth claim.

The intervention analyses complement the path visualizations with quantitative comparisons against random feature ordering. Positive deletion and insertion contrasts support the decision relevance of ranked image regions and text tokens under the evaluated operators. The full-validation text study extends this evidence beyond correctly classified examples, linking the rankings to class-probability changes across the observed population.

5.4. Application Implications

LRIF-Net is intended as the classification component of a review workflow. Posts would enter category-specific queues, where analysts examine the original text and image before accepting a report. Damage-related queues could support checking infrastructure reports, while rescue and donation queues could help organize relief-related information.

The class label provides a retrieval aid, and the visible direct and residual contributions give analysts an additional way to inspect the model. This role is consistent with applications that connect classified posts or extracted damage information to geographic displays [1,2].

The evaluated system remains a decision-support research prototype. A field study should measure category-specific missed reports, analyst review time, and error detection, while connecting predictions to provenance checks and an unknown-category rejection mechanism [40]. Similarity-based rejection would require calibration with held-out unknown categories. Duplicate retrieval, explicit image-text mismatch detection, and object-level damage localization are related extensions, each requiring task-specific targets and evaluation. The ensemble latency and memory measurements provide an initial execution budget for such a study; they do not measure the complete service or the human review process.

5.5. Limitations

Evaluation is limited to the label-consistent CrisisMMD subset; only 17 affected-individuals examples occur in the test split, and the synthetic missing-image, missing-text, and mismatched-image conditions do not characterize naturally incomplete or misleading posts. LOEO provides internal cross-event evidence but not independent-domain validation. Text interventions cover the full validation population, whereas image interventions retain correctly classified subsets; both use one seed. Statistical comparisons condition on fitted models and remain exploratory. Resource measurements cover single-member V100 throughput and three-member latency and allocator memory on one RTX 3060 Laptop GPU; field acquisition, service overhead, energy use, and human review are outside the evaluated scope.

5.6. Future Research Objectives and Plans

Future research will address three objectives. First, it will test LRIF-Net on naturally occurring evidence failures, including missing modalities, blurred or irrelevant images, noisy text, semantic image-text conflicts, and misinformation, rather than only the controlled missing-image, missing-text, and mismatched-image protocols used here. Second, it will evaluate transfer across independently collected crisis events, datasets, regions, and languages, including label systems that differ from the present five-category benchmark and settings with dynamic or previously unseen categories. Third, it will investigate more stable methods for minority-category recognition and quantify how the direct and residual paths, attention interactions, and recurrent updates affect individual decisions. These objectives require new data and prospective experiments; they are not results of the present study.

6. Conclusions

Organizing crisis posts requires preliminary category assignments based on local and global evidence, followed by analyst verification. LRIF-Net addresses the classification stage through the coordinated design of a decision-preserving recurrent residual fusion framework: a complete global-local classifier remains in the final decision, while a compact parameter-tied, zero-initialized branch provides a separately observable additive correction.

On the canonical CrisisMMD split, the fixed three-seed ensemble achieves 91.35% Accuracy, 88.88% Macro F1, and 91.32% Weighted F1. These results indicate competitive performance under the matched evaluation protocol, not predictive superiority: Macro F1 and Balanced Accuracy are lower than for the retained direct classifier, and paired comparisons include zero. The recurrent stage updates 517,509 parameters, reducing final-stage updates by 88.76% and retained non-CLIP parameters by 59.17% relative to matched

standard cross-attention. Gains under missing-image, missing-text, and mismatched-image inputs apply only to the three predefined synthetic conditions. Future work will test naturally occurring missing, noisy, and conflicting multimodal evidence; evaluate transfer across crisis events, datasets, regions, languages, and changing category systems; and develop more stable minority-category methods alongside quantitative analyses of the model's decision and interaction mechanisms. These directions require new data and prospective evaluation and are not presented as current findings.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16188921/s1. The English Supplementary Document provides per-class metrics, detailed ablations, event-level results, frozen intervention protocols, and additional diagnostic evidence.

Author Contributions: Conceptualization, Z.C., Z.P. and S.L.; methodology, Z.C., Z.P. and S.L.; software, Z.C.; validation, Z.C., S.L., Q.L. and H.G.; formal analysis, Z.C.; investigation, Z.C.; resources, Z.P., S.L. and G.L.; data curation, Z.C. and Q.L.; writing—original draft preparation, Z.C.; writing—review and editing, Z.C., Z.P., S.L., Q.L., H.G. and G.L.; visualization, Z.C.; supervision, Z.P. and S.L.; project administration, Z.P. and S.L.; funding acquisition, Z.P. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2024YFC3908000.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of the CrisisMMD data used in this study. CrisisMMD was obtained from its original providers and is available under the dataset's original access conditions; it is not redistributed with this article. The reported aggregate results are contained within the article and Supplementary Materials. Canonical split validation, configuration files, and local preparation instructions are available at https://github.com/ziqic2161-glitch/LRIF-Net (accessed on 4 September 2026).

Acknowledgments: Figure 1 was assembled from openly licensed photographs credited in the figure caption. The authors prepared the illustrative captions and composite layout.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He, C.; Hu, D. Social Media Analytics for Disaster Response: Classification and Geospatial Visualization Framework. *Appl. Sci.* **2025**, *15*, 4330. [CrossRef]
2. Gao, S.; Yang, T.; Xu, Y.; Mou, N.; Wang, X.; Huang, H. Enhancing Disaster Situation Awareness Through Multimodal Social Media Data: Evidence from Typhoon Haikui. *Appl. Sci.* **2025**, *15*, 465. [CrossRef]
3. National Institute of Standards and Technology. DistressNet-NG: Resilient Mobile Broadband Communication and Edge Computing. Available online: https://www.nist.gov/ctl/pscr/distressnet-ng-resilient-mobile-broadband-communication-and-edge-computing (accessed on 28 August 2026).
4. Alam, F.; Ofli, F.; Imran, M. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. *Proc. Int. AAAI Conf. Web Soc. Media* **2018**, *12*, 465–473. [CrossRef]
5. Ofli, F.; Alam, F.; Imran, M. Analysis of Social Media Data Using Multimodal Deep Learning for Disaster Response. *arXiv* **2020**, arXiv:2004.11838.
6. Liu, J.; Singhal, T.; Blessing, L.T.M.; Wood, K.L.; Lim, K.H. CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media, Online, 30 August–2 September 2021; pp. 133–141. [CrossRef]
7. Wang, C.; Nulty, P.; Lillis, D. Transformer-Based Multi-Task Learning for Disaster Tweet Categorisation. In Proceedings of the 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021), Blacksburg, VA, USA, 23–26 May 2021; pp. 705–718.

8. Abavisani, M.; Wu, L.; Hu, S.; Tetreault, J.; Jaimes, A. Multimodal Categorization of Crisis Events in Social Media. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14667–14677. [CrossRef]
9. Cheung, T.H.; Lam, K.M. Crossmodal Bipolar Attention for Multimodal Classification on Social Media. *Neurocomputing* **2022**, *514*, 1–12. [CrossRef]
10. Rezk, M.; Elmadany, N.; Hamad, R.K.; Badran, E.F. Categorizing Crises from Social Media Feeds via Multimodal Channel Attention. *IEEE Access* **2023**, *11*, 72037–72049. [CrossRef]
11. Dar, S.S.; Rehman, M.Z.U.; Bais, K.; Haseeb, M.A.; Kumar, N. A Social Context-Aware Graph-Based Multimodal Attentive Learning Framework for Disaster Content Classification during Emergencies. *Expert Syst. Appl.* **2025**, *259*, 125337. [CrossRef]
12. Li, S.; Liu, Q.; Pan, Z.; Wu, X. CLIP-BCA-Gated: A Dynamic Multimodal Framework for Real-Time Humanitarian Crisis Classification with Bi-Cross-Attention and Adaptive Gating. *Appl. Sci.* **2025**, *15*, 8758. [CrossRef]
13. Wang, Z.; Li, S.; Liu, Q.; Pan, Z.; Sun, X. MKAN-Refine: Fine-Grained Crisis Information Mining via Reliability-Aware Nonlinear Refinement and Kolmogorov–Arnold Networks. *IEEE Access* **2026**, *14*, 66021–66039. [CrossRef]
14. Farah, B.; Cleuziou, G.; Gracianne, C.; Hafiane, A.; Halftermeyer, A.; Canals, R. Image-Text Crisis Tweet Categorization: A Caption-Based Approach. *Proc. Int. ISCRAM Conf.* **2024**, *21*, 1–12. [CrossRef]
15. Munia, N.; Zhu, J.; Nasraoui, O.; Imran, A.-A.-Z. Differential Attention for Multimodal Crisis Event Analysis. *arXiv* **2025**, arXiv:2507.05165. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
17. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6558–6569. [CrossRef] [PubMed]
18. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 457–468. [CrossRef]
19. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114. [CrossRef]
20. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2247–2256. [CrossRef]
21. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep Modular Co-Attention Networks for Visual Question Answering. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6281–6290.
22. Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; Carreira, J. Perceiver: General Perception with Iterative Attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR 139, pp. 4651–4664.
23. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention Bottlenecks for Multimodal Fusion. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 14200–14213.
24. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2022**, arXiv:2106.09685.
25. Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16664–16678. [CrossRef]
26. Lian, D.; Zhou, D.; Feng, J.; Wang, X. Scaling and Shifting Your Features: A New Baseline for Efficient Model Tuning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 109–123. [CrossRef]
27. Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *Int. J. Comput. Vis.* **2024**, *132*, 581–595. [CrossRef]
28. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to Prompt for Vision–Language Models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
29. Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; Li, H. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; pp. 493–510.
30. Khattak, M.U.; Rasheed, H.; Maaz, M.; Khan, S.; Khan, F.S. MaPLe: Multi-Modal Prompt Learning. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19113–19122.

31. Han, Z.; Yang, F.; Huang, J.; Zhang, C.; Yao, J. Multimodal Dynamics: Dynamical Fusion for Trustworthy Multimodal Classification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 20707–20717. [CrossRef]
32. Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J.T.; Peng, X. Provable Dynamic Fusion for Low-Quality Multimodal Data. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; PMLR 202, pp. 41753–41769.
33. Gao, Z.; Jiang, X.; Xu, X.; Shen, F.; Li, Y.; Shen, H.T. Embracing Unimodal Aleatoric Uncertainty for Robust Multimodal Fusion. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 26876–26885.
34. Bi, L.; Zhang, Y.; Wang, L.; Niu, Y.; Zhao, H. Two Challenges, One Solution: Robust Multimodal Learning through Dynamic Modality Recognition and Enhancement. In *Findings of the Association for Computational Linguistics: EMNLP 2025*; Association for Computational Linguistics: Suzhou, China, 2025; pp. 12855–12867. [CrossRef]
35. Peng, X.; Wei, Y.; Deng, A.; Wang, D.; Hu, D. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 8238–8247.
36. Fan, Y.; Xu, W.; Wang, H.; Wang, J.; Guo, S. PMR: Prototypical Modal Rebalance for Multimodal Learning. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 20029–20038. [CrossRef]
37. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1692–1706. [CrossRef]
38. Wang, W.; Tran, D.; Feiszli, M. What Makes Training Multi-Modal Classification Networks Hard? In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12695–12705.
39. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. SMIL: Multimodal Learning with Severely Missing Modality. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2302–2310. [CrossRef]
40. Yu, C.; Hu, B.; Wang, Z. Open-World Disaster Information Identification from Multimodal Social Media. *Complex Intell. Syst.* **2025**, *11*, 7. [CrossRef]
41. Zhang, Y.; Wang, Z.R.; Du, J. Deep Fusion: An Attention-Guided Factorized Bilinear Pooling Network for Audio-Visual Emotion Recognition. *arXiv* **2019**, arXiv:1901.04889. [CrossRef]
42. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 5100–5111. [CrossRef]
43. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 13–23.
44. Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: Universal Image–Text Representation Learning. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; pp. 104–120. [CrossRef]
45. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; Chang, K.-W. What Does BERT with Vision Look At? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5265–5275. [CrossRef]
46. Li, J.; Selvaraju, R.R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
47. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language–Image Pre-Training for Unified Vision–Language Understanding and Generation. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; PMLR 162, pp. 12888–12900.
48. Jain, S.; Wallace, B.C. Attention Is Not Explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Minneapolis, MN, USA, 2–7 June 2019; pp. 3543–3556. [CrossRef]
49. Serrano, S.; Smith, N.A. Is Attention Interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2931–2951. [CrossRef]
50. Petsiuk, V.; Das, A.; Saenko, K. RISE: Randomized Input Sampling for Explanation of Black-Box Models. *arXiv* **2018**, arXiv:1806.07421. [CrossRef]
51. DeYoung, J.; Jain, S.; Rajani, N.F.; Lehman, E.; Xiong, C.; Socher, R.; Wallace, B.C. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4443–4458. [CrossRef]
52. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]

53. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; PMLR 70, pp. 3319–3328.
54. Hooker, S.; Erhan, D.; Kindermans, P.-J.; Kim, B. A Benchmark for Interpretability Methods in Deep Neural Networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9737–9748.
55. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR 139, pp. 8748–8763.
56. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.