Continual Learning for Resilient Multimodal Misinformation Detection Across Sequential Crisis Events

Bilal Ahmed Lodhi, Zeeshan Tariq

School of Computing, Ulster University, York Street, Belfast BT15 1ED, Northern Ireland, UK
b.lodhi@ulster.ac.uk, z.tariq@ulster.ac.uk

Abstract

Misinformation evolves rapidly during crises such as pandemics, political events and natural disasters, challenging the reliability of static detection systems. Although recent multimodal deep learning approaches have integrated text and imagery to improve misinformation detection, they have assumed stationary data distributions and overlooked the sequential domain shifts that are encountered in practice. In this study, we introduce a continual multimodal misinformation benchmark that evaluates resilience across three crisis domains: COVID-19 health misinformation, socio-political misinformation, and disaster-related credibility. We evaluated fine-tuning and replay-based continual learning strategies using a CLIP-based architecture to evaluate their performance. Our results reveal severe catastrophic forgetting under naive fine-tuning, with forgetting exceeding 12%. In contrast, experience replay nearly eliminates forgetting, improves the average task accuracy by over 12 pp, and stabilizes predictive calibration. Our findings establish continual learning as a critical component for building resilient multimodal misinformation detection systems in dynamic real-world environments.

Introduction

The landscape of online misinformation is inherently non-stationary, particularly during crisis periods, when false narratives rapidly emerge, evolve, and cascade across platforms (Starbird et al. 2014), (Vosoughi, Roy, and Aral 2018). From COVID-19 vaccine conspiracies to manipulated disaster imagery during climate events, each crisis domain exhibits distinct linguistic patterns, visual characteristics, and persuasive tactics (Brennen et al. 2020), (Alam et al. 2022). Despite this dynamic reality, current multimodal misinformation detection systems are predominantly trained on static datasets and deployed with the implicit assumption that crisis characteristics remain stable (Zhou and Zafarani 2020), (Kiela et al. 2020).

This static deployment paradigm creates a critical vulnerability: as new crisis events emerge, models must adapt while retaining knowledge of previous misinformation tactics to avoid being misled. As illustrated in Figure 1, real-world misinformation arrives as a non-stationary stream of crisis-specific domains, where models must sequentially adapt to new narratives while preserving knowledge of earlier events through continual multimodal learning. A system trained to detect health misinformation during a pandemic may encounter political deepfakes during election cycles and subsequently face recycled disaster images during climate events. Each adaptation risks catastrophic forgetting, which is the tendency of neural networks to overwrite previously learned knowledge when learning new tasks (Nakamura, Levy, and Wang 2020), (Alam, Ofli, and Imran 2018).

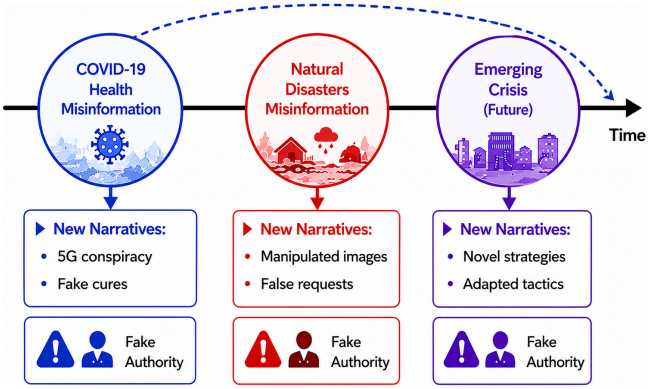

Figure 1: Evolution of crisis-driven misinformation narratives over time. While each crisis introduces new vocabulary and domain-specific content, underlying manipulation strategies persist across events. The solid timeline denotes sequential crisis domains encountered during deployment, while the curved dotted trajectory illustrates the non-linear evolution of misinformation narratives and manipulation strategies.

Recent advances in multimodal deep learning have addressed misinformation detection by jointly modeling textual and visual signals (Alam, Ofli, and Imran 2018)–(Kiela et al. 2020). By leveraging both modalities, these approaches outperform unimodal systems by capturing complementary cues such as manipulated imagery, sensationalist headlines, and misleading contextual framing. However, the prevailing paradigm remains static: models are trained once on a fixed dataset and evaluated under the assumption of an unchanging distribution.

Continual learning offers a framework for addressing this

challenge by enabling models to learn sequentially while preserving the prior knowledge (Chen, Chu, and Subbalakshmi 2021). However, most continual learning research has focused on unimodal vision tasks or reinforcement learning scenarios, with limited exploration of multimodal inputs.

In this study, we bridge this gap by introducing a continual multimodal misinformation detection benchmark that simulates realistic crisis-driven domain shifts. We sequentially trained models across three domains (COVID-19 misinformation, socio-political misinformation, and disaster-related credibility) and evaluated their abilities to retain knowledge, transfer knowledge across domains, and maintain calibrated uncertainty.

We present the first systematic study of multimodal misinformation detection under realistic sequential crisis conditions.

- We constructed a balanced, continual, multimodal benchmark using three real-world datasets representing distinct crisis domains.
- We systematically evaluated fine-tuning and replay-based continual learning in this multimodal setting.
- We analyze not only accuracy, but also catastrophic forgetting, backward transfer, and calibration drift.
- We performed modality ablation to understand how text and visual signals behave during domain evolution.

Our findings challenge the prevailing static deployment paradigm and provide actionable insights for designing resilient misinformation detection systems that can adapt to evolving crisis landscapes while maintaining institutional memory of past misinformation tactics.

Related Work

Early misinformation detection primarily relies on linguistic cues and network propagation patterns. With the proliferation of visually manipulated content, multimodal approaches have become increasingly prominent. Multimodal transformers that jointly embed textual and visual representations have demonstrated substantial improvements over unimodal baselines (Alam, Ofli, and Imran 2018). Large-scale benchmarks such as Fakeddit enable fine-grained multimodal fake news classification (Zhou and Zafarani 2020), whereas cross-modal attention mechanisms and vision–language pre-training further enhance the robustness of detection [6]. Despite these advances, existing methods largely assume static data distributions and do not address evolving misinformation narratives. Catastrophic forgetting remains a fundamental challenge in continual neural learning (Nakamura, Levy, and Wang 2020). Regularization-based approaches, such as Elastic Weight Consolidation, constrain updates to parameters that are critical for previous tasks (Alam, Ofli, and Imran 2018). Replay-based methods preserve knowledge by interleaving stored samples from prior tasks during training (French 1999), with generative replay extending this paradigm through sample synthesis. Although replay has demonstrated strong effectiveness in vision benchmarks, its application in multimodal learning is limited. Prior studies have extensively explored misinformation detection within individual crisis domains (Parisi et al. 2019), (Rebuffi et al. 2017) and continuous learning in vision and language tasks (Riemer et al. 2019), (Lopez-Paz and Ranzato 2017). However, the intersection of these two fields remains largely uncharted. Existing multimodal continual learning research focuses primarily on generic visual reasoning problems (Guo et al. 2017), overlooking domain-specific challenges in crisis informatics, including severe class imbalances, temporal concept drift, and asymmetric costs of misinformation errors. Moreover, modern neural networks are often poorly calibrated, producing overconfident predictions even when incorrect (Rebuffi et al. 2017). This issue is exacerbated by distribution shifts, which raise reliability concerns in high-stakes settings. Although recent studies have begun examining uncertainty in continual learning frameworks (Riemer et al. 2019), calibration dynamics in multimodal misinformation detection remain under-investigated.

Methodology

In real-world deployment, misinformation emerges as a temporally evolving stream of crisis events rather than a single static domain. For instance, a system initially trained on COVID-19 misinformation may subsequently encounter socio-political narratives and later disaster-related misinformation. These domains are observed sequentially, reflecting realistic non-stationary environments. Continual multimodal learning captures this setting by requiring the model to incrementally adapt to each new crisis domain while retaining knowledge of previously encountered misinformation patterns.

We formulated multimodal misinformation detection as a continual learning problem over a sequence of crisis-driven domains. Given a non-stationary stream of multimodal samples consisting of textual and visual content, the model is trained sequentially on tasks corresponding to different crisis contexts without access to the full historical data. Our framework employs a CLIP-based multimodal encoder to extract transferable representations, followed by a fusion network and a classification head for misinformation prediction. To mitigate catastrophic forgetting during domain transitions, we integrated a replay-based continual learning mechanism that interleaves representative samples from prior tasks during optimization. Figure 2 illustrates the overall system architecture and the training flow.

Sequential Crisis Learning Framework

We consider a sequence of tasks T_1, T_2, T_3, each corresponding to a distinct crisis domain. Each task provides multimodal samples $(x_i^{(}(t)), y_i^{(}(t)))$, where $x_i^{(}(t)) = (x_i^{"}text", x_i^{"}image")$ and $y_i \in 0, 1$ indicates misinformation. The model f_θ is trained sequentially:

$$\theta_t = \arg\min_{\theta} E_{(x,y)\sim T_t} \left[\mathcal{L}(f_\theta(x), y)\right] \quad (1)$$

Naive optimization results in forgetting of prior tasks

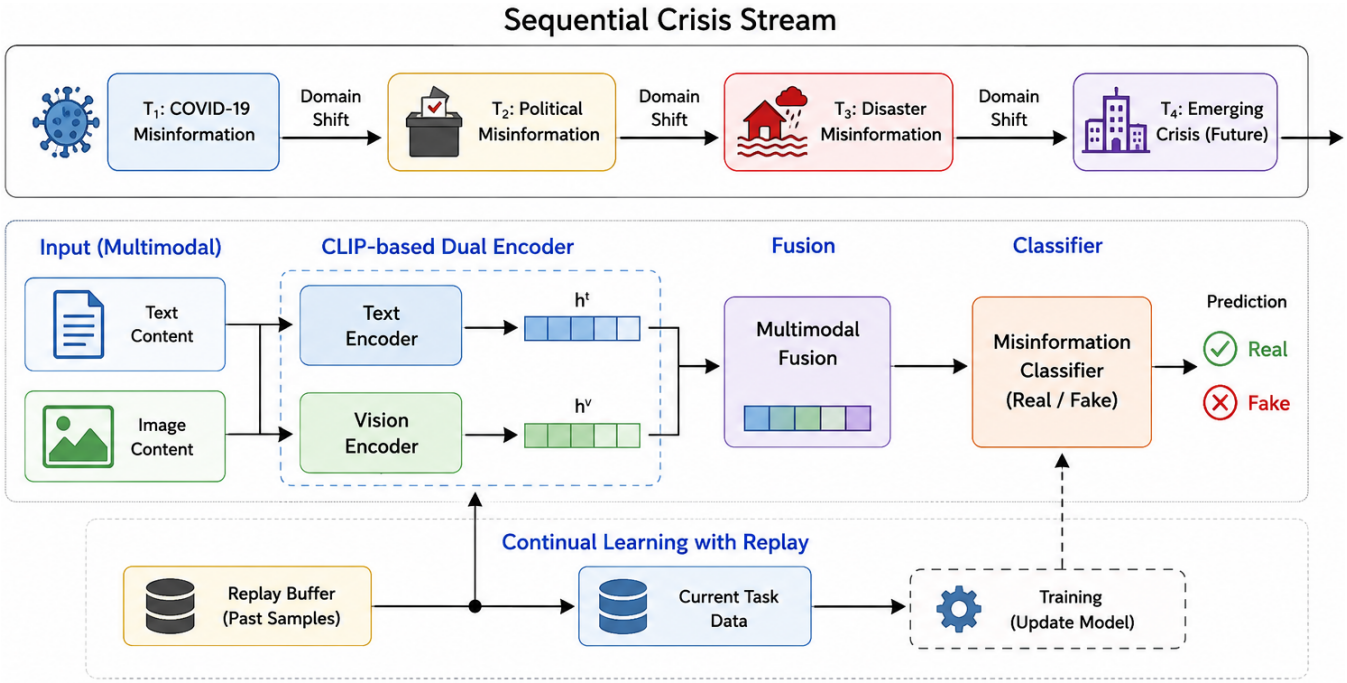

Figure 2: Overview of the continual multimodal learning pipeline with replay-based memory for resilient misinformation detection across crisis domains.

$T_(< t)$. Continual learning seeks to minimize:

$$\sum_{k=1}^{t} E_{(x,y)\sim T_k} \left[\mathcal{L}(f_\theta(x), y)\right] \tag{2}$$

without direct access to complete historical data.

Base Architecture We employed CLIP-ViT (Radford et al., 2021) as the base multimodal encoder because of its strong zero-shot transfer capabilities and demonstrated effectiveness in misinformation detection tasks (Mu et al., 2023). The architecture comprises a vision encoder based on a ViT-B/32 transformer. Multimodal fusion is performed using a late fusion mechanism with learned projection heads that map both modalities into a shared embedding space. The final prediction is computed as

$$f_\theta(x^t, x^v) = \sigma\big(W\left[h^t; h^v; h^t \odot h^v\right]\big) \tag{3}$$

where $h^t = \text{TextEncoder}(x^t)$, $h^v = \text{VisionEncoder}(x^v)$, and $\odot$ denotes element-wise multiplication to capture cross-modal interactions.

Training Protocol

We evaluated both a naive fine-tuning baseline and a replay-based continual learning strategy under a sequential task setting. For the baseline, the model is trained sequentially on the three crisis domains $T_1ßT_2ßT_3$ without any mechanism to retain knowledge from the previous tasks. At each stage, model parameters are updated solely using data from the current task by minimizing the standard binary cross-entropy loss,

$$\mathcal{L} = -\sum_i \left[y_i \log p_i + (1 - y_i) \log(1 - p_i)\right] \tag{4}$$

where y_i denotes the ground-truth label and p_i is the predicted probability. For continual learning, we follow the experience replay strategy of Riemer et al. (2019). A fixed memory buffer Mis stores representative samples from previous tasks, with 500 examples per task selected in a stratified manner to maintain class balance. To enhance representativeness, herding-based selection prioritizing high-loss samples is applied, following Rebuffi et al. (2017). During training on each new task, mini-batches combine 70% current-task samples with 30% replayed samples from the buffer. The total training objective is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{current} + \lambda \mathcal{L}_{replay} \tag{5}$$

with $\lambda = 0.5$ controlling the contribution of replayed data. All models are optimized using a learning rate of $210^(- 5)$, batch size of 32, and weight decay of 0.01. Training is performed for 10 epochs per task, with a dropout rate of 0.1 applied to the fusion and classification layers to mitigate overfitting.

Evaluation Metrics

To comprehensively assess both task-level performance and continual learning behavior, we employed metrics that spanned predictive accuracy, knowledge retention, calibration reliability, and modality contribution.

Continual Learning Metrics To quantify learning dynamics across sequential tasks, we adopt established continual learning measures. Average Accuracy (Diaz-Rodriguez et al., 2018) summarizes overall performance after learning task as

$$A_t = \frac{1}{t}\sum_{i=1}^{t} a_{t,i} \tag{6}$$

where $a_(t, i)$ denotes the accuracy on task iafter training through task t. Forgetting (Chaudhry et al., 2018) captures the extent of performance degradation on previous tasks:

$$F_t = \frac{1}{t-1}\sum_{i=1}^{t-1} \max_{j\in\{1,\ldots,t-1\}} (a_{j,i} - a_{t,i}) \tag{7}$$

measuring the maximum accuracy drop experienced for each past task. Backward Transfer (Lopez-Paz and Ranzato, 2017) evaluates whether learning new tasks improves or degrades prior knowledge:

$$\mathrm{BWT}_t = \frac{1}{t-1}\sum_{i=1}^{t-1}(a_{t,i} - a_{i,i}) \tag{8}$$

where positive values indicate beneficial transfer and negative values reflect interference.

Calibration Metrics To assess the reliability of predicted probabilities, we compute the Expected Calibration Error (ECE) following Guo et al. (2017):

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{9}$$

Expected Calibration Error (ECE) quantifies the mismatch between predicted confidence and empirical accuracy. Predictions are partitioned into M confidence bins B_m. For each bin, $acc(B_m)$ denotes the fraction of correctly classified samples, while $conf(B_m)$ represents the average predicted confidence. ECE computes the weighted average of the absolute differences between these quantities across all bins. Lower ECE values indicate better calibration, meaning that predicted probabilities more accurately reflect true outcome likelihoods. This is particularly important in misinformation detection, where overconfident incorrect predictions can undermine system reliability.

Modality Reliance Metrics To analyze the relative contribution of text and image modalities, we employ gradient-based attribution following (Sundararajan, Taly, and Yan 2017), defined as

$$\mathrm{Attr}_{text} = \frac{\nabla_{h^t} f(x)}{\nabla_{h^t} f(x) + \nabla_{h^v} f(x)} \tag{10}$$

where h^t and h^v represent text and visual embeddings, respectively. Complementarily, we conduct modality ablation studies by removing each modality at inference time and computing the resulting performance drop,which quantifies the dependency of the model on each information source.
The modality attribution score measures the relative contribution of textual and visual features to the model's prediction. It is computed using the gradient magnitudes of the model output with respect to the text and image embeddings. Larger gradient norms indicate greater influence on the prediction. The normalized ratio therefore reflects the relative reliance on each modality, where values closer to 1 indicate stronger dependence on textual information, and lower values indicate greater reliance on visual features. This formulation enables analysis of how modality importance shifts across evolving crisis domains.

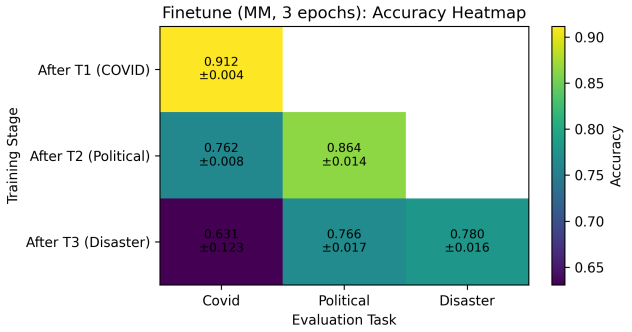

Figure 3: Task-to-task accuracy heatmap under naive finetuning across sequential crisis domains. Rows show the stage after each learned domain, and columns report accuracy per task. Sharp drops on earlier domains reveal strong catastrophic forgetting and negative transfer.

$$\Delta_{modal} = \mathrm{Acc}_{full} - \mathrm{Acc}_{ablated} \tag{11}$$

Results

We evaluated continual multimodal misinformation detection across three sequential crisis domains using naive finetuning and replay-based continual learning strategies. The performance is assessed using the average seen accuracy, catastrophic forgetting, and backward transfer (BWT), capturing both the predictive performance and knowledge retention under domain shifts. Additional analyses were performed to examine the training duration, modality contribution and calibration stability.

Continual Multimodal Performance Across Crisis Domains

Table 1 summarizes the continual learning performance of finetuning and replay-based methods across different training durations.

Naïve fine-tuning suffers from severe catastrophic forgetting, exceeding 12% after three epochs and increasing further with extended training. The consistently negative BWT indicates strong interference between domains, where learning new misinformation narratives degrades previously acquired knowledge. In contrast, experience replay substantially improves continual stability, increasing the average seen accuracy by over 12 pp while nearly eliminating forgetting across training durations. These results demonstrate that replay effectively preserves transferable multimodal representations in the evolving crisis distributions.

Method	Mode	Epochs	Avg Seen Accuracy ↑	Forgetting ↓	BWT ↑
Finetune	MM	3	0.725 ± 0.051	0.126 ± 0.045	-0.189 ± 0.067
Finetune	MM	5	0.693 ± 0.042	0.161 ± 0.047	-0.241 ± 0.071
Replay	MM	3	0.851 ± 0.015	0.002 ± 0.003	0.001 ± 0.007
Replay	MM	5	0.844 ± 0.014	0.007 ± 0.010	-0.003 ± 0.023

Table 1: Continual learning performance across sequential crisis domains and training durations. Results report mean ± standard deviation over three random seeds. Average seen accuracy measures overall task performance, forgetting quantifies performance degradation on previous tasks, and backward transfer (BWT) reflects the influence of new task learning on prior knowledge. Higher accuracy and BWT are better, whereas lower forgetting indicates improved continual resilience.

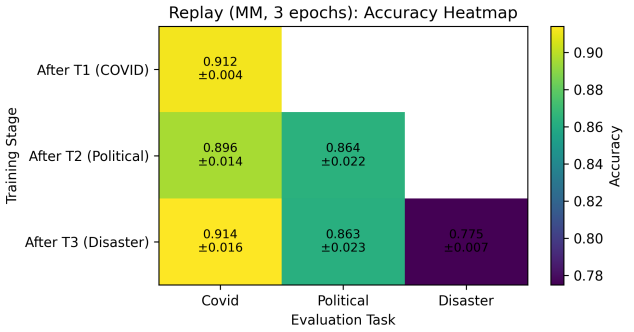

Figure 4: Task-to-task accuracy heatmap under replay-based continual learning. Experience replay preserves performance on previously learned domains while maintaining strong adaptation to new crises, effectively mitigating catastrophic forgetting.

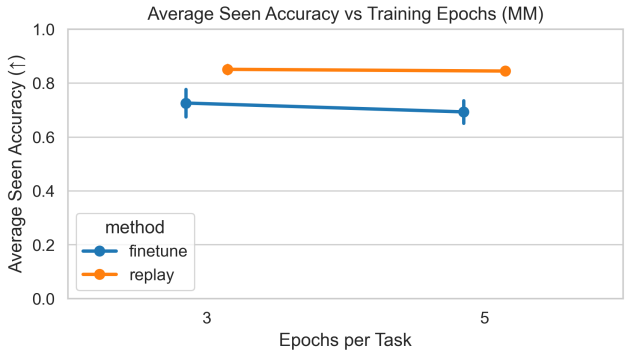

Figure 5: Average seen task accuracy across sequential domains for different training durations. Replay consistently outperforms finetuning and maintains stable performance, indicating improved continual generalization under domain shifts.

The underlying task-to-task accuracy transitions are visualized in Figure 3 and 4 for finetuning and replay-based learning, respectively. Fine-tuning exhibits a pronounced performance collapse in earlier domains following each new task, whereas replay preserves high accuracy across all previously learned tasks.

Impact of Training Duration on Continual Stability

To examine the influence of the optimization length on the continual learning dynamics, we varied the number of training epochs per task and reported the performance in Table 1. Increasing the training duration consistently exacerbates forgetting and negative backward transfer under fine-tuning, indicating an increasing over-specialization to the most recent domain. In contrast, replay-based continual learning maintains near-zero forgetting even with extended training, highlighting its regularizing effect on sequential optimization.

These trends are further illustrated in Figure 5, which plots the average accuracy across training durations, and Figure 6, which depicts forgetting as a function of epochs per task. Although fine-tuning performance deteriorates with prolonged training, replay remains stable, confirming that prolonged optimization alone cannot mitigate domain drift.

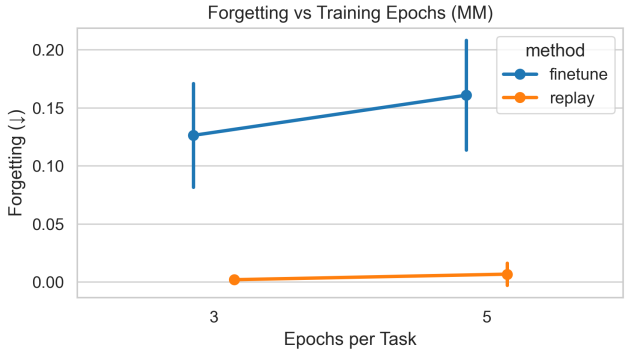

Figure 6: Catastrophic forgetting as a function of training duration per task. Naive finetuning exhibits increasing forgetting with extended optimization, whereas replay-based continual learning maintains near-zero forgetting across training epochs.

Input Modality	Avg Seen Accuracy ↑	Forgetting ↓	BWT ↑
Multimodal (Text + Image)	0.851 ± 0.015	0.002 ± 0.003	0.001 ± 0.007
Text only	0.841 ± 0.017	0.010 ± 0.009	-0.007 ± 0.015
Image only	0.527 ± 0.009	0.103 ± 0.008	-0.154 ± 0.012

Table 2: Modality contribution to continual misinformation detection performance. Replay-based continual learning is evaluated using multimodal, text-only, and image-only inputs to assess cross-domain generalization and robustness under crisis-driven domain shifts.

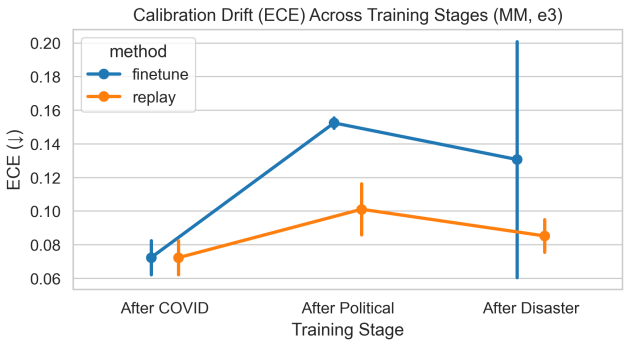

Figure 7: Expected Calibration Error (ECE) across sequential training stages. Naive finetuning leads to increasing miscalibration under domain shifts, whereas replay stabilizes predictive confidence and improves reliability.

Calibration Stability Under Domain Shift

Beyond predictive accuracy, we examined model reliability using the Expected Calibration Error (ECE). Figure 7 shows the calibration behavior across the sequential training stages.

Naive fine-tuning leads to progressively increasing miscalibration as domain shifts accumulate, indicating growing overconfidence in incorrect predictions. Replay-based continual learning substantially stabilizes calibration across tasks, suggesting that preserving historical knowledge improves uncertainty estimation. This indicates that retaining representative samples from prior domains not only mitigates forgetting but also stabilizes confidence estimation under distribution shift. This highlights an important reliability advantage of continual learning strategies for deployment in evolving misinformation environments.

Overall Continual Learning Comparison

Figure 8 provides a consolidated comparison between the fine-tuning and replay strategies across the average seen accuracy, forgetting, and backward transfer. Replay consistently outperformed fine-tuning across all continual learning metrics, achieving higher accuracy, near-zero forgetting, and neutral-to-positive transfer. These results reinforce the effectiveness of replay as a practical mechanism for resilient multimodal learning in non-stationary crisis distributions.

Modality Ablation Under Replay-Based Continual Learning

we analysed the contribution of each modality by evaluating replay based continual learning under multimodal, text only, and image only configurations. The results are shown in Table 2 which reveals a clear modality imbalance in continual generalization. Text only models achieve strong performance with minimal forgetting and approach multimodal accuracies. In contrast, image only models show substantially lower accuracy and severe forgetting, indicating the domain specificity of visual misinformation cues. Although multimodal fusion provides the best overall performance, the dominant contribution of text highlights the limitations of current visual generalization across evolving crises. These trends are further illustrated in Figure 9, showing that while multimodal learning improves overall performance, visual representations remain brittle under narrative shifts, revealing a key limitation of existing multimodal architectures. Although our study covers three representative crisis domains, misinformation also evolves across cultural contexts, platform dynamics, and emerging media formats. Extending continual benchmarks to longer task sequences and finer grained domain shifts remains important future work. While replay effectively mitigates forgetting, its reliance on fixed memory buffers and lack of explicit domain drift or modality stability modelling motivate research into adaptive memory selection and domain aware fusion mechanisms.

Conclusion

In this study, We introduced a continual multimodal benchmark for misinformation detection across evolving crisis domains and showed that static training fails under realistic distribution shifts. Sequential learning on COVID-19, political, and disaster misinformation revealed that naive fine-tuning causes severe forgetting and degraded calibration. In contrast, experience replay greatly improves robustness, preserving accuracy and stable uncertainty. Modality analysis showed that text features generalize more reliably across crises than visual cues, exposing weaknesses in current multimodal fusion and motivating adaptive, domain-aware models. Overall, our results highlight that resilient misinformation detection requires continual learning to adapt to evolving narratives while retaining prior knowledge, and our benchmark provides a foundation for future work on robust multimodal systems in dynamic, high-stakes settings.

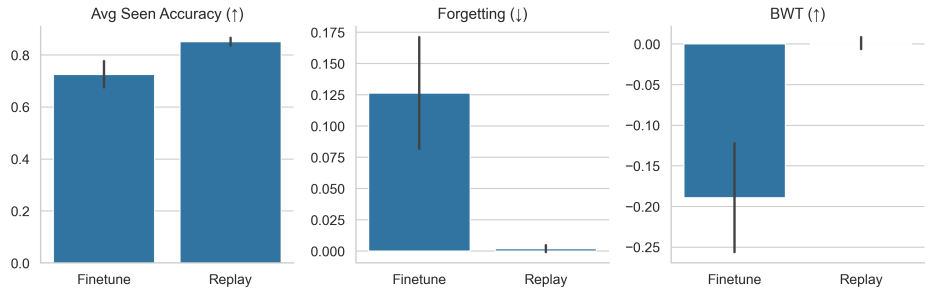

Figure 8: Continual learning performance comparison between finetuning and replay strategies. Bars show mean and standard deviation across random seeds for average seen accuracy, forgetting, and backward transfer (BWT). Replay substantially improves stability and positive knowledge transfer.

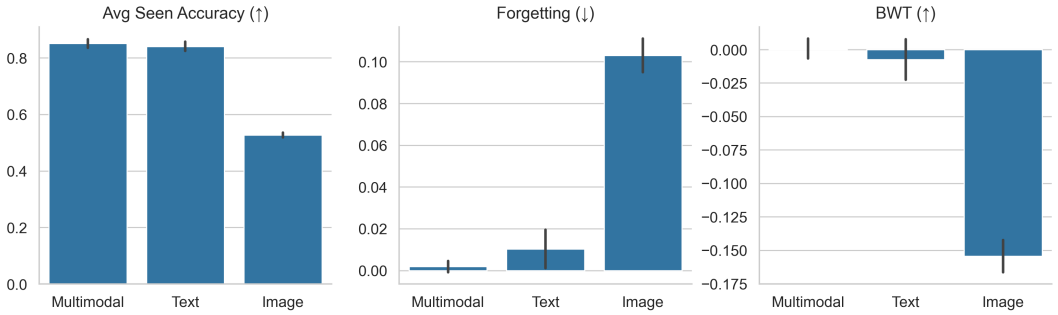

Figure 9: Impact of input modality on continual learning performance under replay-based training. Multimodal fusion achieves the highest overall accuracy and lowest forgetting, while image-only models exhibit poor cross-domain generalization, highlighting modality-specific vulnerabilities.

References

Alam, F.; Ofli, F.; and Imran, M. 2018. CrisisMMD: Multimodal Twitter datasets from natural disasters. In *Proceedings of ICMI*, 465–473.

Alam, F.; et al. 2022. A survey on multimodal disinformation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7815–7835.

Brennen, J. S.; Simon, F.; Howard, P. N.; and Nielsen, R. K. 2020. Types, sources, and claims of COVID-19 misinformation.

Chen, M.; Chu, X.; and Subbalakshmi, K. P. 2021. MMCoVaR: Multimodal COVID-19 vaccine misinformation repository. In *Proceedings of ASONAM*.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of ICML*, 1321–1330.

Kiela, D.; et al. 2020. The Hateful Memes Challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 6467–6476.

Nakamura, K.; Levy, S.; and Wang, W. Y. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of LREC*, 6149–6157.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental classifier and representation learning. In *Proceedings of CVPR*, 2001–2010.

Riemer, M.; et al. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *Proceedings of ICLR*.

Starbird, K.; Maddock, J.; Orand, M.; Achterman, P.; and Mason, R. M. 2014. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon bombing. In *Proceedings of the iConference*, 654–662.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of ICML*, 3319–3328.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Zhou, X.; and Zafarani, R. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5): 1–40.