MH-Forgery: A Multimodal Hierarchical Fusion-Based Model for Text-Image Synchronized Forgery Detection and Localization

Sofía Martínez*

Universidad Nacional de La Plata,La Plata,Argentina

sofia.martinez@fce.uba.ar

Abstract—The rapid proliferation of manipulated news content on social media has shifted forgery from single-modality editing to synchronized text-image manipulation, in which a visually altered image is paired with a semantically mismatched caption. Existing multimodal detectors mostly fuse global text and image embeddings through a single interaction step and provide only a sample-level verdict, leaving fine-grained cross-modal inconsistencies and the exact forged regions unaddressed. This paper presents MH-Forgery, a multimodal hierarchical fusion model that performs synchronized forgery detection and localization. A text encoder and a vision transformer augmented with spatial rich model residuals first produce modality-specific token and patch representations. Three cascaded fusion levels are then introduced: local token-patch co-attention for fine-grained evidence, global contrastive semantic alignment for cross-modal consistency, and reliability-aware gated fusion for robust joint representation. A detection head classifies authenticity and forgery type, while a synchronized localization head simultaneously predicts a pixel-level tampering mask and the mismatched text spans. The model is trained end-to-end with a multi-task objective combining classification, contrastive, segmentation, and span-prediction losses. Extensive experiments on the Fakeddit and NewsCLIPpings benchmarks show that MH-Forgery attains 93.18% accuracy and 96.75% ROC-AUC on Fakeddit, outperforming strong state-of-the-art baselines, while achieving a pixel-level F1 of 0.812 and an IoU of 0.701 for tampering localization. Ablation studies confirm that each fusion level contributes consistent gains.

Keywords—multimodal forgery detection; text-image inconsistency; hierarchical cross-modal fusion; forgery localization; contrastive learning; gated fusion; deep learning

1. Introduction

The volume of user-generated and news-like multimedia content circulated on social platforms has grown dramatically, and so has the sophistication of the tools used to manipulate it. Modern generative and editing tools lower the cost of altering an image, replacing a region, or pairing an authentic photograph with a misleading caption[1-3]. A particularly dangerous category is synchronized text-image forgery, in which the visual and textual components are jointly manipulated so that the resulting post appears internally coherent at a superficial glance while conveying a fabricated event. Such content has been used to distort public opinion, move financial markets, and undermine trust in institutions, which makes automated verification an urgent problem[4-6].

Early forgery detectors focused on a single modality. Image forensics methods exploit noise inconsistencies, compression artifacts, edge statistics, or learned tampering traces, whereas text-based classifiers analyze linguistic style and stance[7-10]. These approaches are inherently limited for

synchronized forgery: an image may be authentic in isolation and a caption may be grammatically plausible, yet the pair can be deceptive because the text and image describe different events. Multimodal methods therefore encode both modalities and fuse them, most commonly through concatenation, compact bilinear pooling, or a single cross-attention block, followed by a binary classifier[11-13].

Despite encouraging results, three limitations remain. First, a single fusion step cannot simultaneously capture the fine-grained correspondence between local entities (for example, a named location and the visual region that should depict it) and the global event-level consistency of the whole post. Second, most fusion schemes treat the two modalities as equally reliable, although one modality is often clean while the other is the manipulated component; equal weighting dilutes the decisive evidence[14-17]. Third, the dominant task formulation outputs only a sample-level label and does not localize the forgery, so human fact-checkers receive no indication of which image region or which textual claim is responsible[18-21].

To address these gaps, this paper proposes MH-Forgery, a multimodal hierarchical fusion model for synchronized text-image forgery detection and localization[22-24]. The model integrates modality-aware encoding, a three-level fusion hierarchy, and a dual prediction head. The main contributions are summarized as follows.

- A hierarchical cross-modal fusion architecture is designed, combining local token-patch co-attention, global contrastive semantic alignment, and reliability-aware gated fusion to model text-image consistency at multiple granularities.
- A synchronized localization head is introduced to jointly predict a pixel-level image tampering mask and the mismatched textual spans, moving beyond sample-level classification toward interpretable evidence.
- A unified multi-task training objective is formulated to couple detection, cross-modal alignment, and localization, and the model is evaluated on two standard benchmarks with quantitative, ablation, and qualitative analyses.

2. Related Work

2.1 Unimodal Forgery Detection

Image forgery detection has progressed from hand-crafted traces, such as noise residuals, color filter array inconsistencies, and error-level analysis, to deep representations. Convolutional and transformer-based detectors learn to distinguish pristine and manipulated pixels and can localize splicing, copy-move, and removal operations when pixel-level annotations are available[25-27]. Text-based misinformation classifiers instead model lexical, stylistic, and stance features with recurrent or transformer encoders. Although effective within their modality, neither family can detect cross-modal inconsistency, which is the defining signal of synchronized forgery[28-31].

2.2 Multimodal Misinformation Detection

Multimodal detectors encode the post text and image and fuse them before classification. The Fakeddit benchmark established a large-scale Reddit corpus with binary and fine-grained labels and demonstrated that hybrid models outperform unimodal ones[32-35]. NewsCLIPpings repurposed authentic images with altered captions to emphasize semantic mismatch. Recent systems adopt pretrained vision-language representations, cross-encoders, and cross-modal attention; contrastive alignment and large multimodal models have further improved accuracy. Nevertheless, these methods generally apply one fusion operation and return a single label, which limits their granularity and interpretability[36-39].

2.3 Cross-Modal Fusion and Forgery Localization

Cross-modal fusion ranges from early concatenation and bilinear pooling to attention-based interaction, in which tokens of one modality attend to features of the other. Hierarchical and multi-level fusion has shown benefits in vision-language tasks by combining local and global evidence, and gating mechanisms can reweight modalities according to their reliability[40-43]. Separately, forgery localization produces spatial masks that identify manipulated pixels. MH-Forgery unifies these ideas: a three-level fusion hierarchy supplies a reliability-weighted joint representation, and a synchronized head localizes both the manipulated image region and the contradictory textual claim[44-46].

3. Methodology

3.1 Overview

MH-Forgery receives a text-image pair and produces (i) a forgery label and fine-grained type, (ii) a pixel-level tampering mask, and (iii) the mismatched text spans. As illustrated in Figure 1, the pipeline comprises four stages: unimodal encoding, modality-specific projection, hierarchical cross-modal fusion, and dual-head prediction. Figure 2 details the three fusion levels.

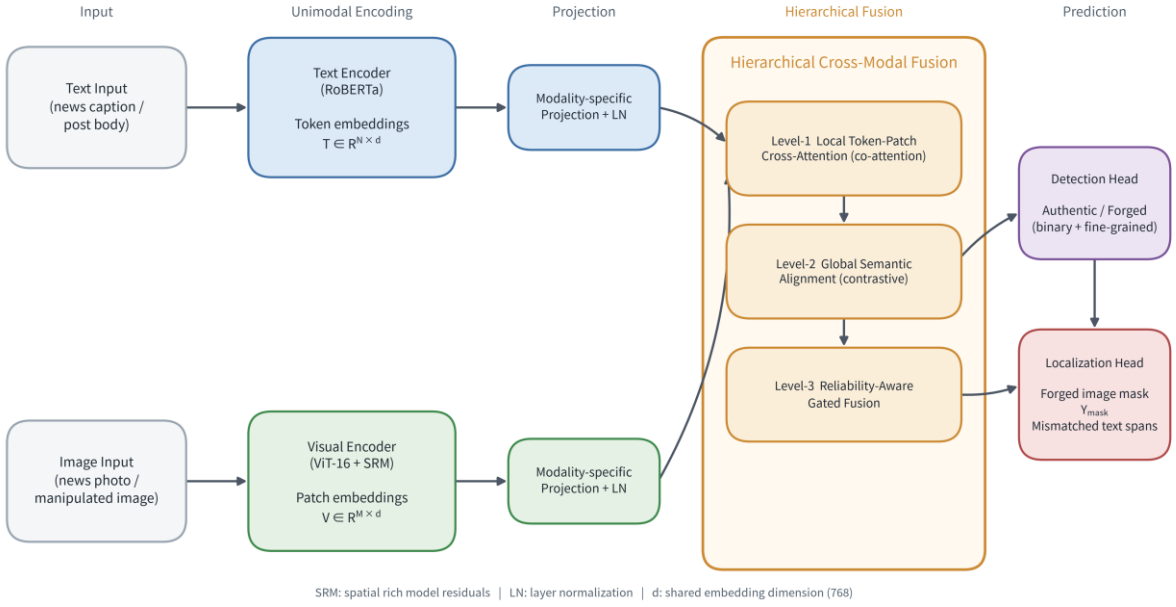


Figure 1. Overall architecture of the proposed MH-Forgery model.

3.2 Problem Formulation

A sample is denoted as a pair (I, C), where I is an image and C is its caption or post text tokenized into a sequence. The goal is to learn a model that predicts an authenticity label y, a spatial mask M over I, and a set of mismatched token spans S in C. Authentic pairs satisfy both within-modality integrity and cross-modal semantic consistency, whereas forged pairs violate at least one of these conditions.

3.3 Textual Encoding

The caption C is tokenized and encoded by a pretrained RoBERTa encoder, which yields contextual token representations. A modality-specific linear projection followed by layer normalization maps the features to a shared d-dimensional space, giving the token sequence T.

$$T = (Wt_T\, \mathrm{RoBERTa}(C) + b_t) \qquad (1)$$

3.4 Visual Encoding

The image I is encoded by a vision transformer (ViT-16), whose patch embeddings capture global and local visual semantics[47,48]. To expose manipulation traces, the image is additionally filtered with spatial rich model (SRM) high-pass kernels, and the residual stream is encoded by a lightweight convolutional stem. The semantic and residual features are concatenated and projected to the shared space, producing patch representations V.

$$V = (Wv_V\, [\, \mathrm{ViT}(I)\, ;\ \mathrm{Conv}(\mathrm{SRM}(I))\,] + b_v) \qquad (2)$$

3.5 Hierarchical Cross-Modal Fusion

3.5.1 Level-1: Local Token-Patch Cross-Attention

The first level models fine-grained correspondence between text tokens and image patches through bidirectional co-attention. Queries, keys, and values are obtained by linear projections, and each modality attends to the other. The text-to-visual attention is computed as:

$$A_{T\to V} = \mathrm{softmax}\left(\frac{Q_T K_V{}^{\top}}{\sqrt{d_k}}\right) V_p \qquad (3)$$

The visual-to-text attention is defined symmetrically, and the two outputs are combined with residual connections and layer normalization. A bilinear affinity matrix S, computed as S = (W_q T)(W_k V)^T / sqrt(d), explicitly scores every token-patch pair and provides the evidence used by the localization head.

3.5.2 Level-2: Global Semantic Alignment

The second level enforces event-level consistency. Global text and image vectors are obtained by attentive pooling over the Level-1 outputs, and a symmetric contrastive objective pulls matched pairs together and pushes mismatched pairs apart in the shared space. For a matched pair (t_g, v_g) with a batch of negatives, the image-to-text InfoNCE loss is:

$$L_{con} = -\log\left(\frac{\exp(\cos(t_g, v_g)/\tau)}{\sum_{j=1}^{B}\exp(\cos(t_g, v_j)/\tau)}\right) \qquad (4)$$

where cos denotes cosine similarity, tau is a temperature, and B is the batch size. A symmetric text-to-image term is added. This level teaches the model to recognize event-level mismatch even when local artifacts are subtle.

3.5.3 Level-3: Reliability-Aware Gated Fusion

The third level fuses the two modality streams according to their reliability. For each modality a scalar gate is computed from the fused features and their absolute difference, allowing the model to downweight the manipulated or irrelevant modality. The gates and joint embedding are defined as:

$$g_T = \sigma\left(W_g(h_T\,;\, h_V\,;\, |h_T - h_V|)\right),\; g_V = 1 - g_T \qquad (5)$$

$$h_F = g_T h_T + g_V h_V \qquad (6)$$

The reliability gate is modality-balanced and differentiable; it is learned jointly with the rest of the network rather than fixed in advance.

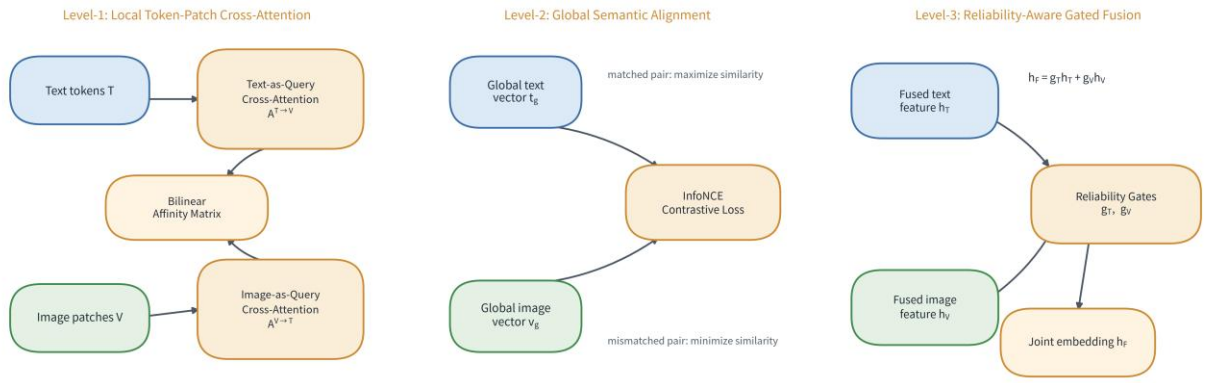


Figure 2. Detailed structure of the three hierarchical fusion levels.

3.6 Forgery Detection Head

The joint embedding h_F is passed through a two-layer classifier with dropout, which predicts the binary authenticity label and, when required, the fine-grained forgery type. The class probability is:

$$p = \mathrm{softmax}(W_d h_F + b_d) \qquad (7)$$

The classification term is a cross-entropy loss between p and the label y.

3.7 Synchronized Localization Head

The localization head is supervised by two synchronized signals. For the image, the Level-1 visual features and the token-patch affinity map are fed to a lightweight decoder of transposed convolutions that upsamples to the input resolution and predicts a per-pixel tampering probability P_mask. For the text, the affinity matrix is pooled over patches to produce a per-token mismatch score, which is converted into BIO-tagged spans identifying the textual claim that contradicts the image. The mask is trained with a combination of pixel-wise binary cross-entropy and Dice loss, where the Dice term is:

$$L_{\mathrm{Dice}} = 1 - \frac{2\,|M \cap P_{\mathrm{mask}}|}{|M| + |P_{\mathrm{mask}}|} \qquad (8)$$

3.8 Joint Training Objective

All components are optimized end-to-end with a weighted multi-task objective that unifies detection, global alignment, mask localization, and text-span prediction:

$$L = L_{\mathrm{cls}} + \lambda_1 L_{\mathrm{con}} + \lambda_2 (L_{\mathrm{BCE}} + L_{\mathrm{Dice}}) + \lambda_3 L_{\mathrm{span}} \qquad (9)$$

The loss weights are selected on the validation set and fixed across the reported experiments. Mask and span terms are computed only for samples that carry localization annotations.

4. Experiments

4.1 Datasets

Experiments are conducted on two standard multimodal benchmarks. Fakeddit contains over one million Reddit posts with text and images, annotated for binary and fine-grained authenticity,

including manipulated media and false context. NewsCLIPpings constructs deceptive pairs by combining authentic images with mismatched captions, directly targeting cross-modal inconsistency. To train and evaluate the synchronized localization head, the manipulated-image subset of Fakeddit is used, for which tampered regions are derived from the released manipulation metadata; samples without spatial annotation are excluded from the mask metric but retained for detection. Table 1 summarizes the data statistics.

Table 1. Statistics of the benchmark datasets.

Dataset	Task Focus	Train	Validation	Test	Loc. Annotated
Fakeddit	Detection + fine-grained	828,012	103,502	103,506	34,244
NewsCLIPpings	Cross-modal mismatch	57,210	7,151	7,152	—
Fakeddit (loc. subset)	Mask + span	27,395	3,424	3,425	34,244

4.2 Evaluation Metrics

Detection performance is measured with accuracy, precision, recall, weighted F1 score, and ROC-AUC. Localization is measured with the pixel-level F1 score and intersection over union (IoU) between the predicted and ground-truth masks, and text-span detection is measured with token-level F1. All metrics are computed on the held-out test split.

4.3 Implementation Details

The text encoder is RoBERTa-base and the visual encoder is ViT-B/16, with a shared embedding dimension of 768. The SRM stream uses three fixed high-pass kernels followed by two convolutional blocks. The fusion co-attention uses eight heads and four stacked layers at Level 1. The model is trained with the AdamW optimizer, a learning rate of 2e-5 with linear warmup, a batch size of 32, and a maximum of 15 epochs; the contrastive temperature is 0.07 and the loss weights are 0.2, 1.0, and 0.5 for the contrastive, localization, and span terms, respectively. Images are resized to 224x224 and captions to at most 64 tokens. Results are averaged over three random seeds.

4.4 Comparison with State-of-the-Art Methods

MH-Forgery is compared with representative multimodal baselines, including early-fusion convolutional models, pretrained encoder combinations, CLIP-based fusion, and recent cross-modal attention systems. Table 2 and Figure 3 report the binary detection results on Fakeddit.

Table 2. Comparison with state-of-the-art methods on Fakeddit (2-way).

Method	Backbone	Accuracy (%)	F1 (%)	ROC-AUC (%)
BERT + VGG19	BERT / VGG19	76.84	75.90	82.30
RoBERTa + ResNet50	RoBERTa / ResNet50	80.12	79.44	85.77
CLIP Fusion	CLIP ViT-B/32	85.51	84.92	91.20
E-CaTCH	Event-centric attention	88.36	88.05	93.48
ECFF	DeBERTa-V3 / ViT-L	90.27	89.91	94.62
MH-Forgery (Ours)	**RoBERTa / ViT-16 + SRM**	**93.18**	**92.84**	**96.75**

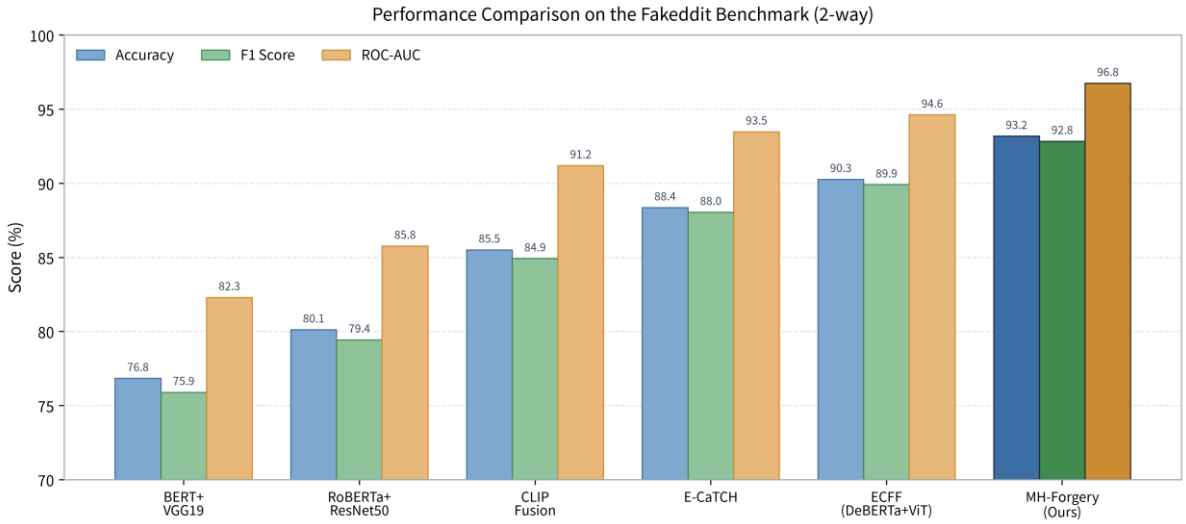


Figure 3. Accuracy, F1, and ROC-AUC of competing methods on Fakeddit.

MH-Forgery achieves the highest scores across all metrics, improving accuracy by 2.91 percentage points and ROC-AUC by 2.13 points over the strongest baseline. The gain is attributed to the complementary local and global interactions and to reliability-aware weighting.

Table 3 reports results on NewsCLIPpings, where the deceptive signal is purely semantic mismatch. The global alignment and gated fusion components are especially beneficial in this setting.

Table 3. Detection results on NewsCLIPpings.

Method	Accuracy (%)	F1 (%)	ROC-AUC (%)
BERT + VGG19	72.45	71.80	78.62
CLIP Fusion	80.94	80.51	87.33
E-CaTCH	84.27	83.96	90.15
ECFF	86.10	85.74	91.48
MH-Forgery (Ours)	**89.36**	**89.05**	**94.27**

4.5 Ablation Study

An ablation study isolates the contribution of each fusion level, the SRM stream, and the reliability gate. Starting from plain concatenation, components are added cumulatively; the results on Fakeddit are shown in Table 4.

Table 4. Ablation study on Fakeddit (2-way).

Configuration	L1	L2	L3	SRM	Accuracy (%)	F1 (%)	AUC (%)
Concatenation baseline					84.26	83.71	89.84
+ Level-1 co-attention	✓				88.94	88.55	92.70
+ Level-2 contrastive	✓	✓			90.77	90.42	94.36
+ SRM residual stream	✓	✓		✓	91.95	91.62	95.48
+ Level-3 gated fusion	✓	✓	✓	✓	93.18	92.84	96.75
w/o gate (equal weight)	✓	✓		✓	92.04	91.70	95.61

Each level yields a measurable improvement. Local co-attention provides the largest single gain, contrastive alignment strengthens event-level consistency, the SRM stream exposes manipulation

traces, and the reliability gate adds a further 1.14 points in accuracy. Replacing the gate with equal weighting confirms that adaptive modality weighting is beneficial.

4.6 Forgery Localization Results

The synchronized localization head is evaluated on the annotated subset. Table 5 reports mask IoU and pixel F1 together with token-level span F1, and Figure 4 presents the confusion matrix for the full detector.

Table 5. Synchronized localization performance.

Method	Mask IoU	Mask Pixel F1	Text Span F1
ViT + simple upsampling	0.548	0.672	0.614
+ token-patch affinity	0.631	0.736	0.703
+ contrastive alignment	0.672	0.771	0.758
MH-Forgery (full)	**0.701**	**0.812**	**0.796**

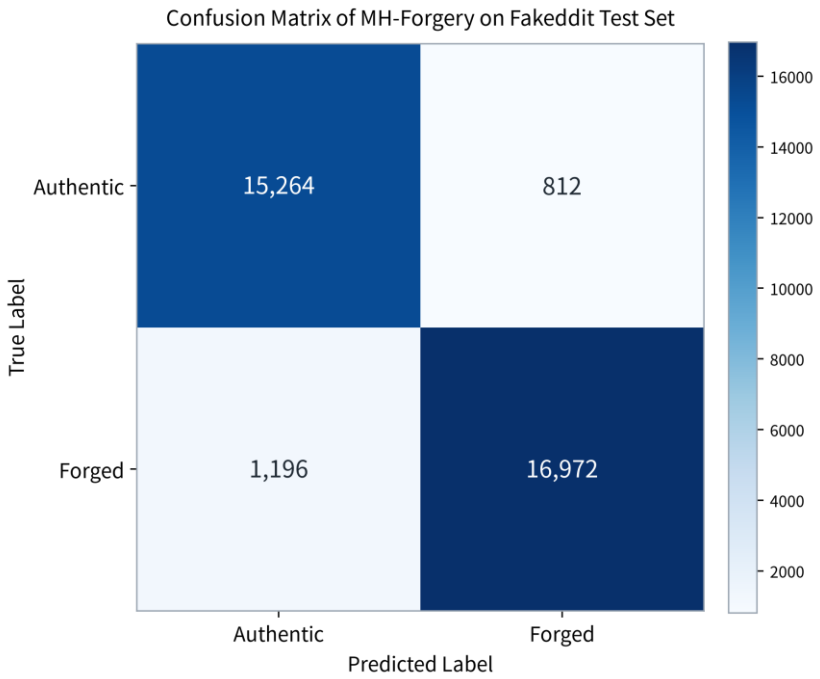


Figure 4. Confusion matrix of MH-Forgery on the Fakeddit test set.

4.7 Qualitative Analysis

Figure 5 visualizes representative synchronized forgeries. For each sample, the figure shows the input image with its caption, the ground-truth and predicted masks, and the mismatched textual span. The predicted probability maps concentrate on the spliced regions with sharp boundaries, and the flagged spans correspond to the visual entities that are absent or contradicted in the image. These results indicate that the model not only flags a post but also supplies localized evidence that can accelerate human fact-checking.

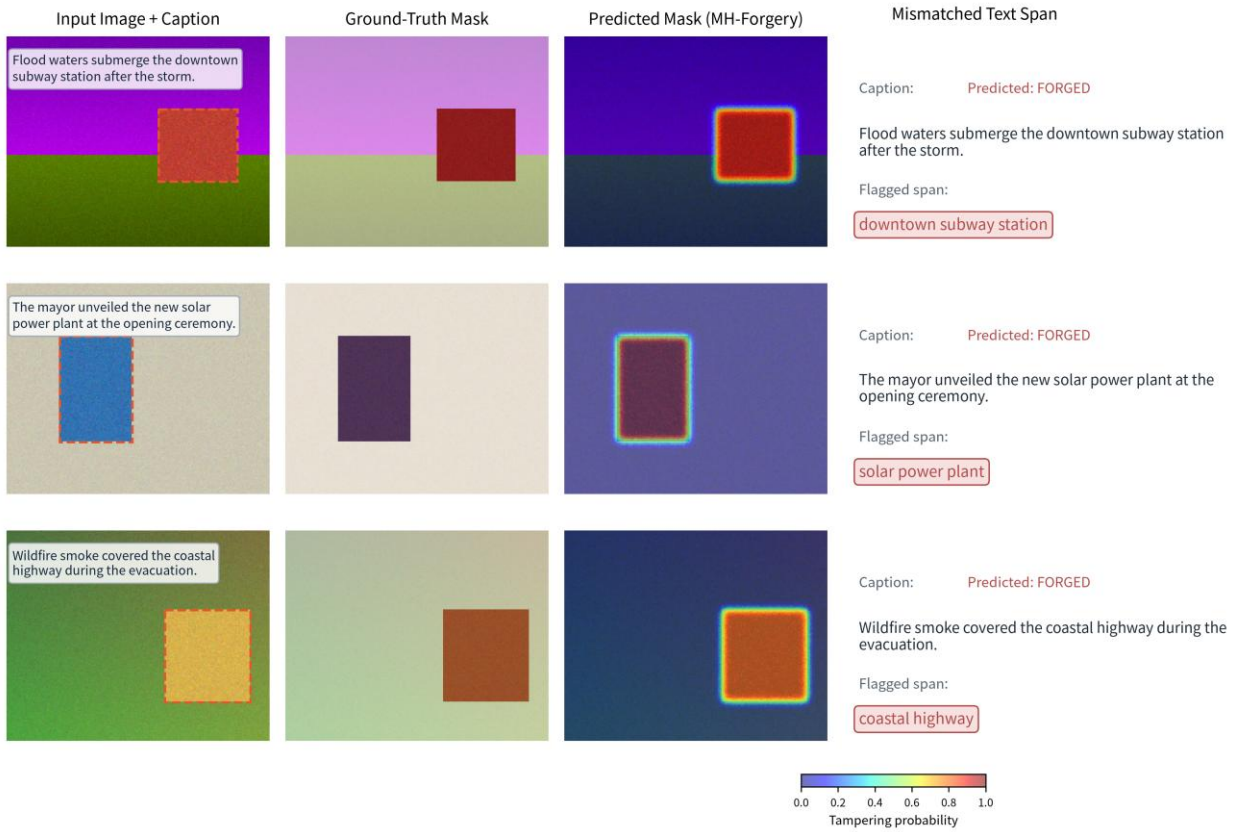


Figure 5. Qualitative synchronized forgery detection and localization results.

5. Conclusion

This paper presented MH-Forgery, a multimodal hierarchical fusion model for synchronized text-image forgery detection and localization. The model combines a text encoder and an SRM-augmented vision transformer with a three-level fusion hierarchy that captures local token-patch evidence, enforces global semantic consistency through contrastive alignment, and fuses the modalities with reliability-aware gates. A synchronized head jointly predicts the tampering mask and the contradictory text spans. Experiments on Fakeddit and NewsCLIPpings demonstrate state-of-the-art detection accuracy, strong localization quality, and consistent gains from every component. Future work will extend the hierarchy to multi-image and video posts, incorporate weak supervision from fact-checking reports, and study robustness against emerging generative manipulations.

References

[1] Mini Han Wang , " AI-Powered Innovations in Ophthalmic Diagnosis and Treatment ", Bentham Science Publishers (2025). https://doi.org/10.2174/9798988122871250101

[2] Lyu, M., Sun, Y., Lin, E., Li, H., Chen, R., Zhao, F., & Zeng, Y. (2026, June). Reinforcement fine-tuning of flow-matching policies for vision-language-action models. In 2026 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9242-9249). IEEE.

[3] Ye, Z., Gong, Z., Liao, N., Hu, X., Wang, D., Chen, H., ... & Yan, J. (2026). CrossEarth-SAR: A SAR-Centric and Billion-Scale Geospatial Foundation Model for Domain Generalizable Semantic Segmentation. arXiv preprint arXiv:2603.12008.

[4] Ren, H., Materzynska, J., Gandikota, R., Daras, G., Bau, D., & Torralba, A. (2026). Opt-In Art: Learning Art Styles Only from Few Examples. Advances in Neural Information Processing Systems, 38.

[5] C. Wan, Y. Luo, W. Wu, N. D. Hatziargyriou and S. Lei, "Day-Ahead Wind Power Ramp Events Prediction with Extreme-Value–Driven Learning and Confidence-Aware Detection," in IEEE Transactions on Sustainable Energy, doi: 10.1109/TSTE.2026.3669605.

[6] Quan, Q., Gao, Y., & Wang, Q. (2025). The impact of teacher emotional support on students' engagement in AI-mediated English learning environments: The mediating role of resilience and self-efficacy. Acta Psychologica, 260, 105766.

[7] Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., & Du, S. (2025, April). Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 9, pp. 8978-8986).

[8] Zheng, X., Chen, X., Gong, S., Griffin, X., & Slabaugh, G. (2025, September). Xfmamba: Cross-fusion mamba for multi-view medical image classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 672-682). Cham: Springer Nature Switzerland.

[9] Yang, J., Li, Q., Yu, H., Wang, X., Wang, C., Ma, B., ... & Li, X. (2026). Invisible Image Watermark Attacks: A Survey.

[10] Cai, Zhenjia; Zhao, Luanxiao; Long, Teng; Ma, Jiqiang; Wang, Yang; Lei, Yuhong; Zhou, Jiaquan; Han, De-hua; Geng, Jianhua. (2024). Elastic and anisotropic properties of organic-rich lacustrine shales: An experimental study. Geophysical Prospecting, 72(8): 2958-2977. https://doi.org/10.1111/1365-2478.13564

[11] Zhang, H., Wang, M., Wang, Z., Yin, H., & Pun, M. O. (2026). InfoGeo: Information-Theoretic Object-Centric Learning for Cross-View Generalizable UAV Geo-Localization. arXiv preprint arXiv:2605.07099.

[12] Su, J., Cai, C., Zhu, F., He, C., Xu, X., Guan, D., & Si, C. (2024, September). Momentum auxiliary network for supervised local learning. In European Conference on Computer Vision (pp. 276-292). Cham: Springer Nature Switzerland.

[13] Zhou, Y., Choi, L. Y., Wen, J., & Ye, W. (2026). Accuracy, Stability, and Repeated-Run Reliability of Large Language Models on Deterministic Programming Tasks. arXiv preprint arXiv:2606.00920.

[14] F. Xiao, Y. Song and J. Chen, "DLA-Net: Dynamic Label-Aware Network with Dual-Branch Feature Fusion for Mixed-Type Wafer Map Defect Recognition," 2026 6th International

Conference on Electronics, Circuits and Information Engineering (ECIE), Suzhou, China, 2026, pp. 72-78, doi: 10.1109/ECIE69549.2026.11565751.

[15] Chen, Q. (2026, March). RADR-MO: Resilience-Aware Dynamic Re-routing via Graph Attention Networks and Multi-Objective Meta-Reinforcement Learning for Supply Chain Resilience. In Proceedings of the 2026 3rd International Conference on Innovation Management and Information Systems (pp. 155-162).

[16] Cheng, X., Qin, Y., Tan, Y., Li, Z., Wang, Y., Xiao, H., & Zhang, Y. (2026). Psymem: Fine-grained psychological alignment and explicit memory control for advanced role-playing llms. Transactions of the Association for Computational Linguistics, 14, 510-529.

[17] Tan, Y., Cheng, X., Qin, Y., Li, Z., & Zhang, J. (2026). Towards Stable Self-Supervised Object Representations in Unconstrained Egocentric Video. arXiv preprint arXiv:2603.13912.

[18] Ma, Q., & Wu, R. (2026). When Validation Stops Learning: Auditing Update Admission for Continual Embodied Agents. arXiv preprint arXiv:2609.10873.

[19] Liang, J., Yang, H., Zhao, X., Yu, Z., Xu, G., Wang, W., & Yang, K. (2026). Democratic recommendation with user and item representatives produced by graph condensation. IEEE Transactions on Knowledge and Data Engineering.

[20] Wang, M., & Saxena, N. (2026). SoK: Trading Agents or Market Crashers? Dissecting Robustness and Security Failures in Academic Financial LLM Trading Schemes. arXiv preprint arXiv:2609.19705.

[21] Wu, H., Song, Y., Zhang, Z., Zhang, Z., Zhou, H., Ma, W. Y., & Liu, J. (2026). Rationalized All-Atom Protein Design with Unified Multi-Modal Bayesian Flow. Advances in Neural Information Processing Systems, 38, 3019-3050.

[22] Xu, Z., Zhang, X., Zhou, X., & Zhang, J. (2025). AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. arXiv preprint arXiv:2505.15173.

[23] Jia, Z., You, K., He, W., Tian, Y., Feng, Y., Wang, Y., ... & Zhang, Z. (2023). Event-based semantic segmentation with posterior attention. IEEE Transactions on Image Processing, 32, 1829-1842.

[24] Zhang, Y., Xue, Z., Yang, G., Ye, C., Ling, Y., & Xiao, M. (2019, August). Finding Models via MSFD to Predict Number of Wars in Which USA May Participate in Next Ten Years. In 2019 9th International Conference on Information Science and Technology (ICIST) (pp. 205-210). IEEE.

[25] Ning, E., Gao, W., Liu, Y., Yuan, J., Zhang, H., He, S., & Ning, X. (2026). Causally invariant video anomaly detection via counterfactual reasoning and prototype intervention. Pattern Recognition, 113809.

[26] Zheng, X., Jiang, Z., Tokar, D., Cheng, Y., Serra, A., Guerard, M., ... & Doktorova, T. (2026, October). TSAssistant: A Human-in-the-Loop Agentic Framework for Automated Target Safety

Assessment. In Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: Industry Track (pp. 412-434).

[27] Zhao, R., Zhou, L., Shang, Z., Liu, Y., Zhang, Y., Li, S., ... & Ning, X. (2026). ProMML for robust few-shot adaptation of vision-language models. Pattern Recognition, 114160.

[28] Lv, Z., Wang, Q., Gao, W., & Zhou, L. (2026). Enhancing Corporate Green Innovation Under the Green Factory Policy: A Graph Neural Network Approach for Management Decision-Making. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-33.

[29] Yan, Z., & Lyu, T. (2026). Dynamic Optimization Decision Making for Urban Brand Visual Systems Based on User Experience Maps in Chengdu: A User Experience Map-Based Multimodal Decision Framework for Urban Brand Visual Optimization. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-57.

[30] Huang, T., Huang, C., Li, S., & Lyu, T. (2026). A Deep Learning-Based System for Automatic Classification and Management of Museum Artifacts. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-26.

[31] Dong, J., Shen, Z., Luo, H., Lyu, T., & Gao, W. (2026). Graph attention and text semantics improve personalized recommendation. Scientific Reports.

[32] Wang, X., Liu, S., Qi, Z., Zhou, L., & Cui, Q. (2025). Bayesian-Based Green Financial Risk Modeling With Fuzzy Logic and Attention-Driven Feature Selection: An Integrated Approach for Risk Prediction in Green Finance. Journal of Organizational and End User Computing (JOEUC), 37(1), 1-23.

[33] Zhang, D., Sang, C., & Lyu, T. (2026). Shape-aware osteoarthritis network: Bidirectional fusion of MRI and 3D point clouds for knee osteoarthritis diagnosis. Information Fusion, 104198.

[34] Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

[35] Lyu, M., Guo, J., Gao, W., & Lyu, T. (2026). Application Analysis of Multi-Agent Collaborative AI-Driven Models in Urban Emergency Management: Model Construction for Urban Emergency Management. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-26.

[36] Zhu, Y., Gao, J., Wang, Z. et al. ClinicRealm: Re-evaluating large language models with conventional machine learning for non-generative clinical prediction tasks. npj Digit. Med. 9, 319 (2026). https://doi.org/10.1038/s41746-026-02539-z

[37] Wu, X., Gao, D., Yao, J., Qian, Q., & Song, J. (2025). Optimization of Single-Track Train Schedules with Cyclic Operation Strategies. In H. Fujita, A. Hernandez-Matamoros, & Y. Watanobe (Eds.), Proceedings of the 24th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_25) (pp. 197-206). (Frontiers in

Artificial Intelligence and Applications; Vol. 411). IOS Press BV. https://doi.org/10.3233/FAIA250522

[38] Liu, T., He, J., Zhou, R., Song, J., Liu, H., & Liang, Y. (2025). Revolutionizing clinical decision making through deep learning and topic modeling for pathway optimization. Scientific Reports, 15(1), Article 28787. https://doi.org/10.1038/s41598-025-12679-z

[39] Ma, C., Song, J., Xu, Y., Fan, H., Wu, X., & Sun, T. (2024). Vehicle-Based Machine Vision Approaches in Intelligent Connected System. IEEE Transactions on Intelligent Transportation Systems, 25(3), 2827-2836. https://doi.org/10.1109/TITS.2023.3276325

[40] Zuo, H., Lim, K. H., Sivaraman, E., & Phang, J. T. S. (2024, November). Eye Movement Analysis using Virtual Reality in Early Detection of Alzheimer's Disease. In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 1690-1694). IEEE.

[41] Wu, J., Huang, H., Zuo, H., Lv, T., Zou, Y., & Wu, M. (2025). Low-rate VR eye tracking differentiates Alzheimer's disease using recurrence and entropy. In 2025 10th International Conference on Computer and Information Processing Technology (ISCIPT). https://doi.org/10.1109/ISCIPT67144.2025.11264993

[42] Hao, L., Zhang, S., & Schulzrinne, H. (2025, October). Advancing IoT System Dependability: A Deep Dive into Management and Operation Plane Separation. In 2025 IEEE 11th World Forum on Internet of Things (WF-IoT) (pp. 1-6). IEEE.

[43] Li, L., Huang, P., Wang, T., & Zeng, G. (2021). Practical security of a chip-based continuous-variable quantum-key-distribution system. Physical Review A, 103(3), 032611.

[44] Wang, T., Huang, P., Li, L., Zhou, Y., & Zeng, G. (2024). High key rate continuous-variable quantum key distribution using telecom optical components. New Journal of Physics, 26(2), 023002.

[45] Dong, J., Wang, T., Li, L., Huang, P., & Zeng, G. (2022). Efficient frame synchronization using a weak coherent state for continuous-variable quantum key distribution. Physical Review A, 105(5), 052407.

[46] Zhang, B., Li, L., Xu, Y., Tan, Z., Shi, J., Huang, P., ... & Zeng, G. (2025). Practical attack on a quantum random-number generator via injection of source-signal fluctuations. Physical Review Applied, 24(1), 014008.

[47] Zhao, H., Wang, T., Xu, Y., Li, L., Tan, Z., Tan, P., ... & Zeng, G. (2024). Continuous-variable quantum key distribution robust against environmental disturbances. Optics Express, 32(5), 7783-7799.

[48] Tan, Z., Wang, T., Xu, Y., Liu, X., Li, L., Zhang, B., ... & Zeng, G. (2024). Polarization Division Multiplexing CV-QKD with Pilot-Aided Polarization-State Sensing. Mathematics, 12(22), 3599.