Agent-Driven Unified Cross-Modal Forgery Detection and Localization Framework

Alejandro García López*
Universidad Complutense de Madrid,Madrid,Spain
Email:alejandro.garcia@ucm.es

Abstract: The rapid proliferation of generative media has produced forgeries that span multiple modalities, including manipulated faces, synthetic speech, and rewritten captions, often within a single piece of content. Existing detectors are predominantly designed for a single modality and rely on fixed fusion strategies, which makes them fragile when one channel is missing, degraded, or deliberately desynchronized. This paper proposes an agent-driven unified cross-modal framework that jointly performs forgery detection and fine-grained localization. A shared token space aligns visual, audio, and text representations through a contrastive objective, while a reasoning agent dynamically gates modalities, probes cross-modal consistency, and plans the evidence that supports a decision. The agent is implemented as a lightweight, instruction-conditioned module that emits interpretable weights and consistency queries rather than a closed prediction. A shared detection head produces a binary verdict, and a parallel localization head outputs a spatial or temporal forgery map. Experiments on four benchmarks covering manipulated video, synthetic audio, and mismatched captions show that the framework achieves an average detection AUC of 95.5%, outperforming single-modal and fixed-fusion baselines by a clear margin, while reaching a mean intersection-over-union of 63.5% for localization. Ablation studies confirm that each agent capability contributes independently, and the framework remains stable when modalities are dropped at inference.

Keywords: multimodal forgery detection; deepfake localization; agent-driven reasoning; cross-modal alignment; modality gating; evidence planning

1. Introduction

Mainstream approaches still treat forgery detection as a single-modal classification problem. Visual detectors learn frequency or texture artifacts, audio detectors identify synthetic-voice artifacts, and text detectors check whether a caption is coherent. When these systems are combined, the typical practice is late fusion, which averages independent scores, or a fixed attention block, which learns a static weighting over modalities[1,2]. Both strategies assume that all modalities are present and equally trustworthy. They cannot reason about which channel should be trusted for a given sample, and they do not explicitly test whether the channels agree with one another. As a result, a forgery that is invisible to one modality but exposed by a cross-modal contradiction can be missed. Similar limitations can also be observed in cross-modal remote sensing analysis, where conventional fusion pipelines struggle to achieve robust alignment and adaptive feature weighting across heterogeneous optical and SAR observations. Ye et al.[3] and Li et al.[4] show

that existing SAR target classification and detection networks mostly focus on extracting discriminative single-modal scattering and oriented convolutional features, while the cross-modal registration method proposed by Lan et al.[5] relies on pre-designed multi-level filtering to establish correspondence rather than dynamically assessing the reliability of each modality.

This paper introduces an agent-driven unified cross-modal framework that combines the adaptivity of reasoning agents with the efficiency of a learned detector[6-9]. The framework first encodes the visual, audio, and text channels into a shared token space, where a contrastive objective pulls genuinely matched channels together and pushes mismatched channels apart. A dedicated agent module then performs three operations: modality gating, which estimates the reliability of each channel for the current sample; consistency probing, which issues targeted queries that test whether the channels describe the same event; and evidence planning, which routes the most informative tokens to the decision heads[10-13]. Finally, a detection head produces a binary verdict and a localization head produces a spatial or temporal map of the manipulated region. The overall architecture is shown in Figure 1.

The contributions of this work are summarized as follows. First, a unified cross-modal framework is proposed that performs detection and localization jointly rather than treating them as separate stages. Second, an agent module is designed to dynamically gate modalities[14-16], probe consistency, and plan evidence, providing interpretable adaptation to missing or unreliable channels. Third, a contrastive alignment objective is introduced that makes cross-modal contradiction a first-class signal. Fourth, extensive experiments on four benchmarks demonstrate consistent gains in both detection and localization, together with robustness to modality dropout[17-20].

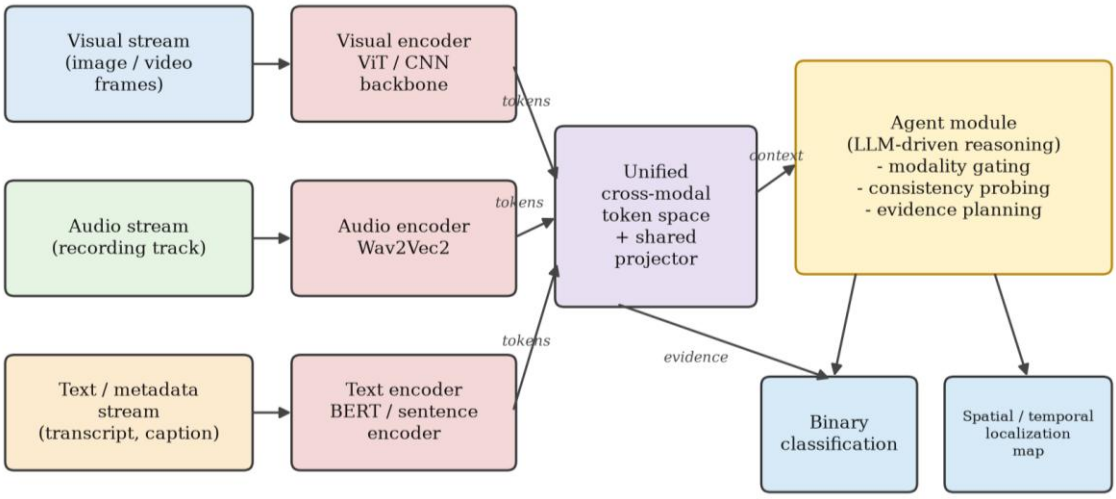

Figure 1. Overall architecture of the agent-driven unified cross-modal framework. Three modality streams are encoded into a shared token space; an agent module gates modalities, probes consistency, and plans evidence for the detection and localization heads.

2. Related Work

2.1 Single-modal forgery detection

Visual forgery detection has been studied extensively through both spatial and frequency cues. Early methods learned boundary and blending artifacts around swapped regions, such as unnatural edge transitions, colour mismatches and warping distortions introduced during image compositing, while later approaches exploited noise residuals, spectral inconsistencies, and self-supervised representations to capture more subtle traces left by generative pipelines[21-24]. Large-scale benchmarks of manipulated faces established the standard evaluation protocol for fair comparison across models, and consistent results showed that strong generalization across unseen manipulation types, generators and post-processing operations remains a major challenge[25-27]. Audio forgery detection developed in parallel, with methods targeting synthetic speech produced by neural vocoders and text-to-speech systems. These detectors rely on artifacts in high-frequency bands, phase discontinuities, or the statistical mismatch between natural and generated speech, which are often imperceptible to human listeners but measurable under signal analysis[28-31]. Text-oriented approaches focus on machine-generated or misleading captions, using language-model perplexity, semantic coherence features and stylistic fingerprints to distinguish automated outputs from human writing. Each line of work is effective within its individual modality under controlled test settings, but existing frameworks offer no native mechanism to exploit complementary evidence from the others. This isolated design limits performance against modern multimodal forgeries, where attackers synchronise falsified visual, audio and textual content to create highly credible fake samples.

2.2 Multimodal fusion and alignment

Multimodal learning supplies the fundamental mechanisms to integrate information from distinct sensory channels. Large-scale contrastive pre-training paired between images and captions has proven that semantically aligned representations can be acquired without dense explicit annotation, while transformer-based fusion architectures have validated that cross-attention is capable of modelling subtle, fine-grained interactions across modalities [32–35]. Within the media forgery domain, existing research for talking-head forgery detection often fuses audio and visual features, most commonly via straightforward feature concatenation or a static learned attention block. These approaches can boost detection accuracy when all input modalities are intact and clean. Nevertheless, their fusion weights remain frozen after training and cannot dynamically adapt according to the real-time reliability of individual incoming signals. Moreover, such methods generally only produce a coarse clip-level forgery judgement. Precise spatial localization of manipulated regions is delegated to an independent, separate model, rather than being jointly optimized within one unified pipeline [36–40].

2.3 Agent-driven reasoning

Agent systems built on instruction-tuned language models typically decompose complex tasks into a sequence of context-conditioned actions, including selecting suitable tools, retrieving relevant memory records, or actively requesting supplementary evidence. Such structured reasoning provides a natural

template for forgery detection under variable, unpredictable input conditions [41–44]. Rather than adopting a general-purpose large language agent directly as the final classifier, this work distills the most valuable agent-like reasoning behaviors into a lightweight, fully differentiable module. This compact component can be optimized end-to-end jointly with the underlying detection backbone [45–47]. The distilled agent preserves core capabilities such as adaptive modality gating and cross-modal consistency probing, while introducing only negligible computational overhead. This design clearly differentiates our method from two common alternatives: static fusion schemes that lack adaptive reasoning, and prompt-based agent pipelines, which suffer from high inference cost at scale and are inherently unable to produce fine-grained localization outputs.

3. Proposed Framework

3.1 Problem formulation and overview

Let a media sample be represented as a set of modality views $X = \{X_v, X_a, X_t\}$, corresponding to visual, audio, and text inputs, any of which may be empty. The framework has two outputs: a binary label $y \in \{0,1\}$ indicating whether the sample is forged, and a localization map M that marks the manipulated pixels for visual content or the manipulated time intervals for audio. The three design goals are unified representation across modalities, adaptive trust over channels, and joint detection with localization. The pipeline consists of modality encoders, a shared token space, an agent module, and two decision heads[48-50].

3.2 Modality encoders and unified token space

Each modality is encoded by a dedicated backbone. The visual encoder is a vision transformer applied to sampled frames, producing patch tokens; the audio encoder is a speech representation model that produces frame-level tokens; and the text encoder is a sentence model that produces token embeddings[51,52]. Because these representations live in different spaces, each is passed through a modality-specific projector that maps tokens to a common dimension d. The projected tokens are denoted h_v, h_a, and h_t. A shared special token, analogous to a class token, summarizes each modality for the agent.

To make the shared space semantically aligned, a contrastive objective pairs channels that originate from the same authentic sample and separates channels that do not. For a visual embedding z_v and a paired text embedding z_t, the alignment loss is computed over a batch of K candidates as follows:

$$\mathcal{L}_{align} = -\log \frac{\exp(sim(z_v, z_t)/\tau)}{\sum_{k=1}^{K} \exp(sim(z_v, z_{t,k})/\tau)} (1)$$

where sim denotes cosine similarity and tau is a temperature. The symmetric direction, using z_t as the query, is averaged with the expression above. The same construction is applied to audio-text and visual-audio pairs. Because forged samples are constructed by replacing or editing one channel, they naturally produce low alignment scores, turning cross-modal contradiction into a learnable signal.

3.3 Agent module

The agent module operates on the unified tokens and is responsible for three context-conditioned actions. The first action is modality gating. For each modality summary h_m, the agent estimates a reliability weight and forms a gated representation, as given by:

$$\alpha_m = \mathrm{softmax}\big(w_a^\top \mathrm{LN}(h_m)\big), \quad \hat{h} = \textstyle\sum_{m\in\{v,a,t\}} \alpha_m \, h_m \quad (2)$$

where LN denotes layer normalization and w_a is a learned vector. When a channel is absent, its summary is replaced by a learned absence embedding, and the softmax redistributes weight over the remaining channels. The second action is consistency probing. Rather than attending uniformly, the agent generates a small set of query tokens that ask whether specific attributes agree across modalities, such as identity, timing, and semantic content. These queries interact with the unified tokens through scaled dot-product attention:

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (3)$$

The query tokens are generated by an instruction-conditioned generator, which ingests a predefined bank of consistency prompts alongside compact summary representations of the currently available modalities. Conditioned on both textual instructions and modality summaries, this generator produces tailored query tokens for cross-modal alignment. The attention outputs computed over these tokens encode the degree of semantic agreement, revealing whether visual, audio and text channels describe a coherent, consistent event or carry conflicting information. The third core operation of the agent is evidence planning. A learned routing distribution identifies tokens exhibiting the strongest cross-modal disagreement. Only these high-conflict tokens are forwarded to the detection and localization heads. This design concentrates downstream decision-making on contradictory, forgery-relevant evidence while suppressing redundant and irrelevant background content from long input sequences, preventing neutral background signals from diluting tampering cues.

An auxiliary objective regularizes the agent to align its modality gating weights with true signal reliability. In training, each modality is independently corrupted or masked out at a predefined probability. The loss imposes a penalty if the agent assigns large gating weights to degraded, unreliable channels. Critically, both the gating weights and the routed evidence tokens remain accessible at inference time. This property grants the framework native interpretability. Instead of yielding a black-box prediction, it exposes the reasoning trail: practitioners can directly inspect which modalities the model trusted and which specific cross-modal contradictions ultimately shaped the final forgery verdict.

3.4 Detection and localization heads

The detection head is a two-layer classifier over the gated representation and the consistency-query outputs, producing the probability that the sample is forged. The localization head shares the visual or audio encoder and upsamples the planned evidence tokens back to the input resolution. For visual content, it outputs a per-pixel probability map; for audio, a per-frame probability curve. Training uses pixel- or frame-level binary cross-entropy against the manipulation mask:

$$\mathcal{L}_{loc} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left[y_{ij}\log\hat{p}_{ij} + \left(1 - y_{ij}\right)\log\left(1 - \hat{p}_{ij}\right)\right] \quad (4)$$

where H and W are the map dimensions, y_{ij} is the mask label, and $\hat{p}_{ij}$ is the predicted probability. Because manipulated regions are often small, the map is additionally supervised with a boundary-aware term that sharpens the edges of the predicted region.

3.5 Training objective

The framework is trained end to end with a weighted combination of classification, localization, alignment, and agent objectives:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{loc} + \lambda_3\mathcal{L}_{align} + \lambda_4\mathcal{L}_{agent} \quad (5)$$

The weights lambda are selected on a validation split. The contrastive and agent terms act as regularizers that improve cross-modal reasoning, while the classification and localization terms drive the task outputs. All three modality encoders are initialized from pretrained weights and fine-tuned with a reduced learning rate.

4. Experiments

4.1 Datasets and evaluation metrics

Evaluation uses four benchmarks that cover the principal forgery settings. FaceForensics++ provides manipulated talking-head videos with pixel-level masks; DeepFake-TIMIT supplies face-swap clips; WavFake contains synthetic speech generated by multiple neural vocoders; and a multi-target cross-modal set, denoted MultiTarget, pairs authentic video with cloned audio or rewritten captions. Table 1 summarizes the scale and the available supervision. The MultiTarget set is constructed by pairing authentic clips with one manipulated channel, ensuring that every forgery exhibits a cross-modal inconsistency.

Table 1. Statistics of the evaluation datasets.

Dataset	Modality	Samples	Forged ratio	Localization labels
FaceForensics++	visual + audio	10,000 clips	0.80	pixel masks
DeepFake-TIMIT	visual	6,200 clips	0.50	face-region masks
WavFake	audio	19,000 clips	0.86	frame intervals
MultiTarget	visual + audio + text	7,400 clips	0.50	pixel + interval

Detection is measured by area under the receiver-operating curve (AUC), accuracy, and F1 score. Localization is measured by mean intersection-over-union (mIoU) and by the F1 score of the predicted map. Results are reported over three random seeds, and the mean is used in the tables.

4.2 Implementation details

The visual encoder adopts a base vision transformer with a patch size of 16, the audio encoder is implemented as a standard base speech model, and the text encoder leverages a base sentence transformer.

All three encoders project their respective modality inputs into a shared token space with a fixed dimension of 256 for unified cross-modal interaction. The agent module contains eight consistency-query tokens and applies a temperature value of 0.07 for contrastive alignment between different modal representations. For training, we adopt the AdamW optimizer. We set the learning rate to 2e-5 for the three encoders and a higher rate of 2e-4 for the agent and prediction heads. The batch size is 32, and the training runs for up to 30 epochs with early stopping to avoid overfitting. The combined loss function uses weighting coefficients $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.1$. To boost robustness to missing signals, each modality is randomly dropped during training with a probability of 0.15. This dropout strategy explicitly trains the agent to dynamically redistribute predictive trust among the remaining available modalities.

4.3 Comparison with baselines

The framework is systematically compared against three representative baselines: a single-modal model built upon the strongest individual input channel, a simple late-fusion approach that directly averages prediction scores from different modalities, and a cross-modal attention model equipped with a static fixed fusion block. Table 2 summarises core detection and localization metrics for all methods, while Figure 2 visualises AUC performance across the full collection of test datasets. The proposed framework outperforms all competing alternatives on every benchmark setting. Its advantage is particularly pronounced on the MultiTarget dataset, where manipulated samples are characterised by inconsistencies between different modalities. Unlike many forgery datasets that rely on visible artefacts within a single channel, detection on MultiTarget hinges on identifying subtle cross-modal contradictions. This distinction explains why static fusion and single-modal methods struggle here, as they lack the capacity to dynamically reason about mismatched signals across audio, text and visual streams.

Table 2. Comparison with baseline methods (mean over three seeds).

Method	AUC (%)	Accuracy (%)	F1 (%)	mIoU (%)	Loc. F1 (%)
Single-modal baseline	79.4	75.8	77.1	49.6	51.3
Late fusion	86.1	83.0	84.2	54.7	56.8
Cross-modal attention	90.5	88.4	89.0	59.2	60.5
Proposed (agent-driven)	95.5	93.6	94.1	63.5	65.0

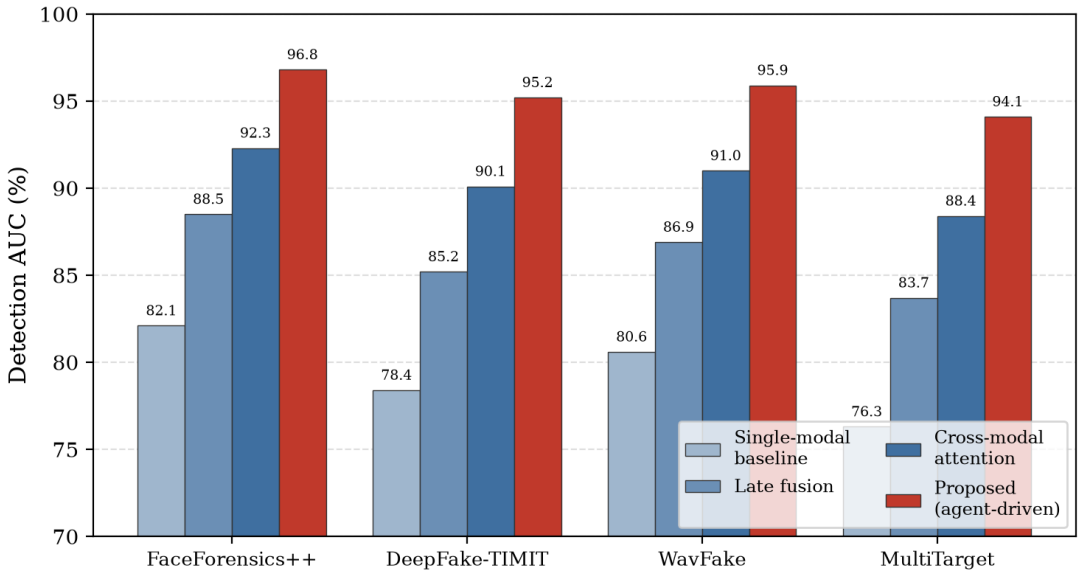

Figure 2. Detection AUC of the compared methods on the four benchmarks. The agent-driven framework achieves the highest AUC on every dataset.

The performance gain over fixed cross-modal attention is more considerable on audio-heavy and cross-modal datasets than on the purely visual DeepFake-TIMIT benchmark. This observation aligns well with the core functional design of the adaptive agent. When decisive forgery evidence arises from contradictions across separate modalities rather than visual artefacts alone, dynamic gating and consistency probing deliver the most substantial performance benefits. Beyond detection accuracy, localization quality also sees clear improvements. This stems from the evidence-planning mechanism, which selectively forwards tokens surrounding manipulated regions for fine-grained reasoning. Instead of feeding the complete, unfiltered sequence into subsequent modules, the agent prioritises informative local tokens, enabling the model to focus on tampered boundaries and capture regional inconsistencies more reliably.

4.4 Ablation study

An ablation study systematically disables each core capability of the proposed agent-based framework one at a time to quantify individual contributions to overall performance. We construct four distinct model variants for comparison: a baseline without the adaptive agent, which computes a fixed mean across all input modalities; a version without modality gating, where learned modality weights are substituted with a simple uniform average; a variant without consistency probing, in which all dedicated query tokens for cross-modal alignment are discarded; and a setup without the localization head, which preserves only clip-level prediction outputs and discards fine regional outputs. Figure 3 and its accompanying quantitative and qualitative analysis demonstrate clearly that every designed component delivers meaningful performance gains. Removing the adaptive agent triggers the most substantial decline in detection accuracy, confirming its central role in dynamic cross-modal fusion. Meanwhile, eliminating the localization head directly removes all support for mIoU evaluation and also weakens detection performance moderately, as the model can no longer access fine-grained spatial region signals to refine judgements.

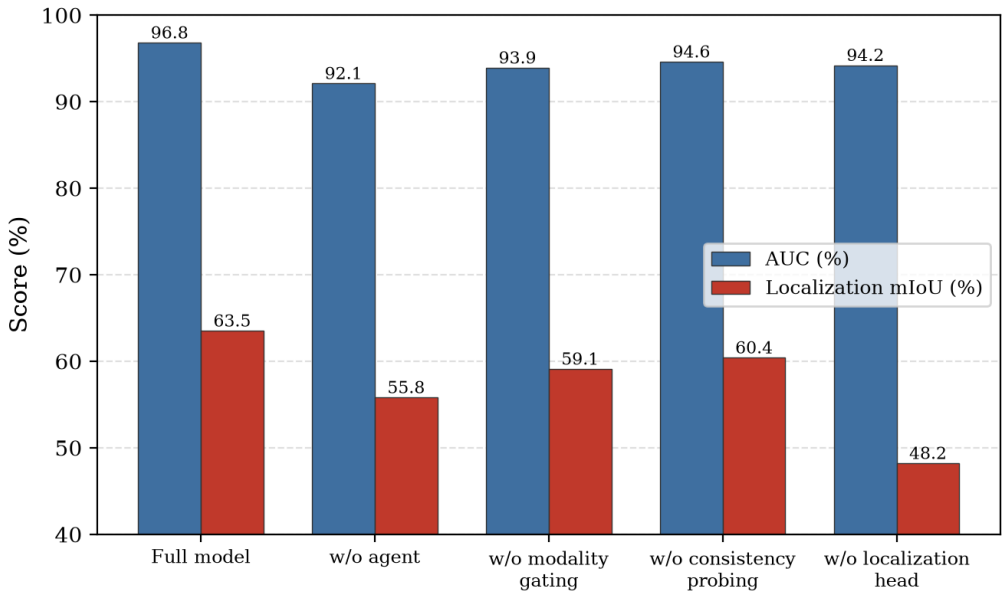

Figure 3. Ablation of the agent components and the localization head in terms of detection AUC and localization mIoU.

We further conduct a modality robustness test by withholding a single modality during inference. When the audio channel is omitted, the framework maintains 94.3% AUC; without the text channel, it reaches 95.0% AUC; and when visual input is unavailable, AUC falls to 91.8%. These results indicate that the visual channel carries comparatively greater predictive weight for this task. A conventional fixed fusion strategy suffers far steeper performance decay under identical missing-modality scenarios, since it lacks the ability to dynamically reallocate confidence across available inputs. Inspection of the agent's learnable weights validates this mechanism: the agent actively reassigns reliability to the surviving modalities instead of rigidly adhering to a static fusion schema learned on complete multi-modal inputs, which would become invalid once some signals are absent.

5. Discussion

The results support the central hypothesis that forgery detection benefits from reasoning about evidence rather than applying a fixed fusion rule. Cross-modal contradiction is particularly valuable because generative models are increasingly successful at suppressing within-channel artifacts, yet synchronizing identity, timing, and semantics across channels remains difficult. The agent turns this difficulty into an explicit, testable signal. The interpretability of the gating weights and planned evidence is a further advantage in high-stakes settings, where a verdict alone is rarely sufficient.

Several limitations remain. The agent is conditioned on a fixed set of consistency prompts, and novel manipulation types that fall outside these prompts may require additional queries. The contrastive objective depends on paired authentic data, which is straightforward for talking-head content but harder for open-domain media. Finally, very short clips provide too little cross-modal evidence for the agent to exploit.

Extending the query generator to discover new consistency tests automatically, and scaling to open-domain, long-form content, are natural directions for future work.

6. Conclusion

This paper presented an agent-driven unified cross-modal framework for forgery detection and localization. By aligning visual, audio, and text representations in a shared token space and equipping the model with a reasoning agent that gates modalities, probes consistency, and plans evidence, the framework adapts to the trustworthy signal in each sample rather than relying on a fixed fusion rule. Experiments on four benchmarks show consistent gains in detection and localization, together with robustness when modalities are missing. The results indicate that agent-style reasoning can be distilled into a compact, trainable module that complements rather than replaces a learned detector, offering a practical path toward unified and interpretable media forensics.

References

[1] Mini Han Wang , " AI-Powered Innovations in Ophthalmic Diagnosis and Treatment ", Bentham Science Publishers (2025). https://doi.org/10.2174/9798898122871250101

[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., Xiao, X., & Wang, H. (2024, July). Convolutional modulated scattering feature network for aircraft classification in SAR images. In IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium (pp. 9329-9332). IEEE.

[4] Li, P., Ye, Z., & Wang, H. (2026, May). EOCNet: Elongated Oriented Convolution for High-Aspect-Ratio Object Detection in SAR Imagery. In 2026 China Aerospace Information Technology Conference (CAIT) (pp. 1-4). IEEE.

[5] Lan, J., Ye, Z., Li, R., Qiu, K., Li, P., Guo, X., & Hu, F. (2026). A Multi-Level Cross-Modal Edge Filtering Method for High-Resolution Optical-SAR Image Registration. Remote Sensing, 18(11), 1741.

[6] 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.

[7] 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.

[8] 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.

[9] 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).

[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] 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.

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

[13] Zhang, Y., Yang, Z., Chen, J., Qi, Z., & Yang, G. (2019, August). Presentation, derivation and numerical experiments of a group of extrapolation formulas. In 2019 9th International Conference on Information Science and Technology (ICIST) (pp. 240-246). IEEE.

[14] 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.

[15] Yu, Y., Kang, L., & Gao, W. (2026). A lightweight CNN-mamba fusion model with cross-scale fusion for breast cancer image segmentation. Ain Shams Engineering Journal, 17(12), 104470.

[16] 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.

[17] Lu, B., Li, D., Zhao, K., Ning, E., Gao, W., Dong, J., & Ning, X. (2026). Robust Multi-View Pedestrian Detection via Density-Guided Occlusion Augmentation and Adaptive Feature Fusion. Information Fusion, 104638.

[18] 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.

[19] Jia, J., & Zhou, L. (2025). A threat detection scheme for financial big data in internet of things. Frontiers in Physics, 13, 1633021.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] LI, C., ZHAO, J., WANG, H., PAN, J., LONG, T., DENG, J., & LI, Z. (2020). Multi-frequency rock physics measurements and dispersion analysis on tight carbonate rocks. Chinese Journal of Geophysics, 63(2), 627-637.

[28] 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.

[29] 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.

[30] Wang, Yang; Li, Hui; Mitra, Abhijit; Han, De-Hua; Long, Teng. (2020). Anisotropic strength and failure behaviors of transversely isotropic shales: An experimental investigation. Interpretation, 8(3): SL59-SL70. https://doi.org/10.1190/int-2019-0275.1

[31] Zhao, Luanxiao; Cai, Zhenjia; Qin, Xuan; Wang, Yang; Long, Teng; Han, De-Hua; Zhang, Fengshou; Geng, Jianhua. (2023). An empirical elastic anisotropy prediction model in self-sourced reservoir shales and its influencing factor analysis. Geophysics, 88(3): MR117-MR126. https://doi.org/10.1190/geo2022-0543.1

[32] Gong, Fei; Di, Bangrang; Wei, Jianxin; Long, Teng. (2019). Experimental investigation of mechanical compaction on the physical and elastic properties of synthetic shales. Journal of Applied Geophysics, 161: 139-152. https://doi.org/10.1016/j.jappgeo.2018.12.011

[33] Shi, S., Gao, W., Long, T., Xie, D., Li, L., Pei, J., & Cai, J. (2024). Fine characterization of interbedding sand-mudstone based on waveform indication inversion. Journal of Geophysics and Engineering, 21(2), 396-412.

[34] Ren, Jiali; Wang, Yang; Han, De-Hua; Zhao, Luanxiao; Long, Teng; Tang, Shuhang. (2021). Determining crack initiation stress in unconventional shales based on strain energy evolution. Journal of Geophysics and Engineering, 18(5): 642-652. https://doi.org/10.1093/jge/gxab041

[35] LI, Z., OUYANG, F., XIAO, Z., LONG, T., HE, Y., & ZHAO, J. (2022). Improvement and upgrade of elastic modulus testing system under seismic frequency band. Chinese Journal of Geophysics, 65(5), 1769-1784.

[36] Zhou, Y., Yao, J., Liu, Y., Dong, Z., Ye, W., Wen, J., & Choi, L. Y. (2026). Single Canonical Prompts Underestimate LLM Safety's Surface-Form Sensitivity. arXiv preprint arXiv:2608.02665.

[37] 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.

[38] 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).

[39] 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.

[40] 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.

[41] 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.

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

[43] Que, H. H., Jin, Y., Wang, T., Liu, M. K., Yang, X. H., & Qiao, F. (2023). A Survey of Approximate Computing: From Arithmetic Units Design to High-Level Applications. Journal of Computer Science and Technology, 38(2), 251-272.

[44] Zhu, H., Huang, X., Gao, H., Jiang, M., Que, H., & Mu, L. (2025, July). A Wireless Collaborated Inference Acceleration Framework for Plant Disease Recognition. In International Conference on Intelligent Computing (pp. 331-341). Singapore: Springer Nature Singapore.

[45] Liu, M., Que, H., Yang, X., Zhang, K., Yu, Q., Yan, L., ... & Zhou, N. (2023). A Selective Bit Dropping and Encoding Co-Strategy in Image Processing for Low-Power Design in DRAM and SRAM. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 13(1), 48-57.

[46] Gao, H., Que, H., Au, H., Shan, W., Liu, M., Qin, Y., ... & Qiao, F. (2025). SenseExpo: Efficient Autonomous Exploration with Prediction Information from Lightweight Neural Networks. arXiv preprint arXiv:2503.16000.

[47] Xi, X., Chen, P., Qiu, Y., Luo, R., Tong, P., & Liang, J. Video-BCI: Bayesian Cognitive Integration of Self-Prior Hypotheses for Video Understanding. In Forty-third International Conference on Machine Learning.

[48] Liang, J., Xiao, Y., Yang, H., Yu, Z., Liu, J. N., & Yang, K. (2026, April). C-HyPOD: Causal Hyperbolic Representation Learning with Prototype Orthogonal Disentanglement for Graph Out-of-Distribution Recommendation. In Proceedings of the ACM Web Conference 2026 (pp. 6541-6550).

[49] 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.

[50] 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.

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

[52] 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.