GEF-VA: Gated Evidence Fusion for General AI-Generated Video-Audio Detection

Shuaibo Li[1†], **Laixin Zhang**[2†], **Wei Ma**[2], **Weilin Ruan**[1], **Hongqiu Wang**[1], **Jialu Li**[1], **Lei Zhu**[1*]

[1]The Hong Kong University of Science and Technology (Guangzhou)

[2]Beijing University of Technology

sli270@connect.hkust-gz.edu.cn, leizhu@hkust-gz.edu.cn

Abstract

Recent advances in generative models have enabled highly realistic and synchronized video-audio synthesis, significantly increasing the risk of multimodal forgery. However, existing detectors largely treat multimodal forgery as a binary classification problem, overlooking the underlying configuration structure where manipulations may occur in the visual stream, the audio stream, or their cross-modal interaction. This leads to configuration-agnostic representations that limit discrimination across diverse forgery patterns. In this work, we propose **G**ated **E**vidence **F**usion for General AI-Generated **V**ideo-**A**udio Detection (GEF-VA), a configuration-aware framework that models multimodal forgery as structured evidence reasoning. GEF-VA first preserves modality-specific forensic cues via foundation embedding sets extracted from frozen visual and audio encoders, and then transforms them into forgery-aware representations through lightweight adapters. To enable adaptive reasoning, a cross-modal gating mechanism dynamically allocates evidence according to input-dependent modality reliability, producing a unified gated representation for prediction. Furthermore, we introduce a two-stage training strategy, where the model is first trained on the full dataset to learn general evidence structures, followed by hard-sample refinement to enhance robustness under challenging configurations. Experiments on both the validation set and the official DDL 2.0 challenge verify the effectiveness of GEF-VA, which achieves a Binary Score of 92.88 and a four-class score of 78.51 on the challenge.

1 Introduction

Recent advances in generative models have dramatically improved the realism and modality diversity of synthetic media [Li *et al.*, 2026a; Zhang *et al.*, 2025a]. Beyond synthesizing high-quality video and audio content, modern generative systems further enable cross-modal generation and direct production of highly synchronized video-audio conten. While this progress unlocks broad creative value, it also significantly amplifies the risk of realistic misinformation and multimedia manipulation. As generative capabilities continue to evolve, forgery patterns become increasingly heterogeneous, spanning manipulations in the visual stream, the audio stream [Borsos *et al.*, 2023], and their cross-modal composition [Hu *et al.*, 2025; Khalid *et al.*, 2021b]. This shift fundamentally complicates the detection problem, as reliable judgment can no longer be achieved through isolated modality analysis [Miao *et al.*, 2025].

A key challenge arises from the fact that multimodal forgery is inherently configuration-based rather than binary. A sample may contain forged video with real audio, real video with forged audio, or inconsistencies between otherwise plausible streams. However, most existing detection paradigms still reduce this structured space into a single authenticity prediction [Khalid *et al.*, 2021b; Li *et al.*, 2026b]. This motivates a fundamental question: how should a detector represent and reason over heterogeneous forensic evidence under different modality configurations?

Early forgery detectors mainly focus on a single modality [Chen *et al.*, 2024; Li *et al.*, 2025; Li *et al.*, 2021]. Visual detectors identify manipulations by modeling spatial artifacts [Ma *et al.*, 2025; Li *et al.*, 2024], temporal inconsistencies [Song *et al.*, 2024], or pretrained foundation representations [Chang *et al.*, 2024], while audio detectors capture acoustic anomalies in generated speech or sound [Xie *et al.*, 2026]. Although effective within their respective domains, these methods inherently overlook cross-modal structure. As a result, they are insufficient for multimodal scenarios where forgery may exist in only one stream or emerge from cross-modal inconsistency.

Recent studies have explored multimodal forgery detection from two perspectives. One line of work combines visual and audio predictions via ensemble or late-fusion strategies [Khalid *et al.*, 2021a; Ilyas *et al.*, 2023], treating each modality independently and merging decisions at the output level. Another line introduces cross-modal consistency modeling, such as phoneme-viseme mismatch [Agarwal *et al.*, 2020], video-audio correspondence learning [Zou *et al.*, 2024], and multimodal contrastive objectives [Yang *et al.*, 2023]. Despite these advances, most existing methods are still trained under binary supervision or coarse-grained consistency signals. Consequently, visual artifacts, acoustic

traces, and cross-modal discrepancies are compressed into a single fused representation, which lacks explicit structure to disentangle modality-specific evidence from cross-modal interaction. This limits the model's ability to generalize across different forgery configurations.

To address the limitations of configuration-agnostic fusion, we introduce **G**ated **E**vidence **F**usion for General AI-Generated **V**ideo-**A**udio Detection (GEF-VA), a configuration-aware evidence reasoning framework that models multimodal forgery detection as a structured evidence organization problem. At the core of GEF-VA is a multi-granularity evidence learning mechanism grounded in the diverse authenticity configurations of video-audio media. Instead of directly compressing visual and audio streams into a single fused feature, GEF-VA first preserves modality-specific forensic cues through foundation embedding sets. Frozen visual and audio foundation encoders provide complementary evidence carriers, while lightweight forgery adapters transform these embeddings into visual and audio forgery representations. This modality-preserving stage ensures that visual artifacts and acoustic traces remain explicitly available before cross-modal reasoning. Building on these representations, GEF-VA introduces an adaptive cross-modal gate that estimates sample-dependent modality reliability and produces a gated video-audio embedding. This mechanism enables the model to dynamically allocate evidence according to different forgery configurations, where the decisive cues may arise from the visual stream, the audio stream, or their cross-modal relation.

As shown in Figure 1, GEF-VA comprises three tightly integrated components and a two-stage training strategy. First, *Visual Representation Adaptation* converts the visual foundation embedding set into a visual forgery embedding, preserving multi-level visual evidence from global semantics to local generation traces. Second, *Audio Representation Adaptation* maps the audio foundation embedding set into an audio forgery embedding, capturing acoustic evidence from the audio stream. Third, *Multi-granularity Authenticity Prediction* performs adaptive gating and jointly optimizes modality-level, video-audio configuration-level, and overall authenticity objectives under multi-granularity supervision. To further improve robustness to challenging cases, GEF-VA is trained in two stages. The first stage learns a general evidence structure from the full training set, enabling the adapters, gate, and predictor to capture broad modality-specific and cross-modal patterns. The second stage performs hard-sample refinement, fine-tuning the trainable modules on difficult samples where evidence is ambiguous or configuration-sensitive. This training scheme first establishes a stable global detector and then strengthens its discriminative ability on challenging video-audio forgery patterns, leading to more reliable configuration-aware prediction.

Our contributions are summarized as follows:

- We formulate multimodal forgery detection as a configuration-aware evidence reasoning problem, and propose GEF-VA, a unified framework that explicitly models modality-specific evidence and cross-modal interactions under a multi-granularity supervision structure.
- We introduce a modality-preserving representation learning strategy and an adaptive cross-modal gating mechanism, enabling dynamic evidence allocation across visual and audio streams for robust forgery detection.
- Extensive experiments on the DDL 2.0: AIGC Safety from Dual Perspectives of Model and Content challenge demonstrate the effectiveness of our method, achieving a Binary Score of 92.88 and a four-class score of 78.51.

2 Related Work

2.1 AI-Generated Video Detection

Early AI-generated video detection methods mainly focus on face-centric deepfakes. These approaches exploit facial priors such as landmark dynamics, head pose inconsistency, physiological signals, spatial artifacts, or temporal coherence to distinguish manipulated facial videos from real ones [Li *et al.*, 2018; Gu *et al.*, 2021]. Later studies further improve facial forgery detection with stronger spatio-temporal networks, data augmentation, or multi-view facial representations [Xu *et al.*, 2023; Peng *et al.*, 2024]. Although effective on facial deepfake benchmarks, these methods rely heavily on face-specific cues. With the rapid development of text-to-video and general video generation models, recent works shift toward open-domain AI-generated video detection. Some methods enhance temporal modeling with optical flow, motion residuals, or spatio-temporal anomaly cues [Song *et al.*, 2024; Bai *et al.*, 2024], while others adopt Transformer- or Mamba-based architectures to capture long-range temporal dependencies and spatial generation artifacts [Kundu *et al.*, 2025; Chen *et al.*, 2024]. Recent studies also explore physics-aware constraints, second-order temporal statistics, and foundation-model representations to improve robustness under unseen generators [Zhang *et al.*, 2025b; Zheng *et al.*, 2025].

2.2 AI-Generated Audio Detection

Recent AI-generated audio detection methods mainly adopt a two-stage pipeline consisting of a pretrained front-end feature extractor and a task-specific back-end classifier. Self-supervised and foundation audio models, such as Wav2Vec [Schneider *et al.*, 2019], XLS-R [Babu *et al.*, 2022], and MMS [Pratap *et al.*, 2024], are widely used to extract robust acoustic representations from large-scale pretraining data. Based on these representations, recent studies further design task-specific back-end classifiers, such as Mamba-based sequence models [Gu and Dao, 2024], to better capture spoofing cues. These works show that pretrained audio representations are effective for improving audio-only forgery detection under unseen or out-of-domain conditions. Beyond detector architectures, recent works further explore how to exploit foundation representations more effectively. Some methods aggregate embeddings from different Transformer layers, select spoof-sensitive layers, or attentively merge hidden states to capture discriminative cues at multiple abstraction levels [Martín-Doñas and Álvarez, 2022; Zhang *et al.*, 2024]. Although these approaches advance single-modality forgery detection, they remain limited to isolated evidence. In

video-audio media, reliable detection requires reasoning over both modality-specific cues and their authenticity configuration, motivating our configuration-aware multimodal framework.

2.3 AI-Generated Video-Audio Detection

With the rapid development of multimodal generation, AI-generated video-audio detection has received increasing attention. Early video-audio forgery benchmarks such as FakeAVCeleb [Khalid *et al.*, 2021b] introduce different combinations of manipulated visual and audio streams, motivating detectors to move beyond unimodal authenticity prediction. Some methods adopt ensemble or late-fusion strategies, combining separately produced visual and audio predictions [Khalid *et al.*, 2021a]. However, such designs often treat the two modalities independently, leaving the relationship between visual and acoustic content insufficiently modeled. Recent studies further explore video-audio consistency modeling and joint representation learning [Mittal *et al.*, 2020; Chugh *et al.*, 2020; Oorloff *et al.*, 2024]. These methods improve multimodal reasoning by learning correspondences between visual and acoustic streams or by designing fusion mechanisms that integrate their features. Despite these advances, most methods are still developed for human-face or speech-based scenarios, and are optimized with binary labels or coarse consistency supervision [Miao *et al.*, 2025]. As a result, visual artifacts, acoustic traces, and cross-modal discrepancies are often merged into a single representation without explicitly reflecting the underlying video-audio authenticity configuration. In contrast, GEF-VA targets general AI-generated video-audio detection by preserving modality-specific evidence and performing configuration-aware fusion.

3 Preliminaries

We follow the DDL 2.0: AIGC Safety from Dual Perspectives of Model and Content Track 2 setting for general AIGC audio-video detection. The benchmark is based on MVAD [Hu *et al.*, 2025], which contains paired video-audio samples with diverse visual styles, content categories, generation methods, and video-audio forgery types. Each input sample consists of a visual stream and an audio stream:

$$x = (I_v, I_a), \tag{1}$$

where I_v denotes the video input and I_a denotes the audio input. For each sample, the detector is required to predict two labels. The first is a binary label:

$$y_b \in \{0, 1\}, \tag{2}$$

where $y_b = 0$ denotes a real sample and $y_b = 1$ denotes a forged sample. The second is a four-class label:

$$y_c \in \{0, 1, 2, 3\}, \tag{3}$$

where the label definition follows:

$$y_c = \begin{cases} 0, & \text{Real-Real(RR)}, \\ 1, & \text{Fake-Fake(FF)}, \\ 2, & \text{Fake-Real(FR)}, \\ 3, & \text{Real-Fake(RF)}. \end{cases} \tag{4}$$

From the four-class configuration label, we further derive two modality-level auxiliary labels for training:

$$y_v = \begin{cases} 1, & y_c \in \{1, 2\}, \\ 0, & y_c \in \{0, 3\}, \end{cases} \qquad y_a = \begin{cases} 1, & y_c \in \{1, 3\}, \\ 0, & y_c \in \{0, 2\}. \end{cases} \tag{5}$$

Here, y_v and y_a indicate whether the video and audio streams are real or fake, respectively. The binary label is also determined by the four-class configuration label:

$$y_b = \begin{cases} 0, & y_c = 0, \\ 1, & y_c \in \{1, 2, 3\}. \end{cases} \tag{6}$$

Thus, for each input x, the model outputs both a binary prediction and a four-class prediction:

$$f(x) = (\hat{y}_b, \hat{y}_c). \tag{7}$$

This setting requires the detector to identify whether a sample is forged and which video-audio authenticity type it belongs to.

4 Method

We introduce GEF-VA, a configuration-aware evidence learning framework for general AI-generated video-audio detection. GEF-VA consists of three core modules: Visual Representation Adaptation, Audio Representation Adaptation, and Multi-granularity Authenticity Prediction. Specifically, the visual and audio branches first extract foundation embedding sets from the input streams using frozen foundation encoders, and then transform these embeddings into modality-specific forgery representations through lightweight adapters. Based on these representations, Multi-granularity Authenticity Prediction performs adaptive cross-modal gating and produces multiple authenticity predictions. To further improve model performance, we adopt a two-stage training strategy, which first learns general evidence structures from the full dataset and then refines the model on hard samples. The overall architecture is shown in Figure 1.

4.1 Visual Representation Adaptation

Visual evidence in generated videos is often distributed across multiple levels, from global scene coherence to frame-level temporal variation and local appearance irregularities. Directly compressing the visual stream into a single descriptor may discard complementary cues before the detector learns which evidence is useful for forgery reasoning. To preserve such information, we first sample frames from the input video I_v and feed them into a frozen visual foundation encoder $\mathcal{E}_v$, obtaining a set of visual foundation embeddings for downstream adaptation.

For each sampled frame, we construct a compact visual token set rather than using only the classification token:

$$R_v^t = \left[r_{\text{cls}}^t; r_{\text{reg}}^t; \bar{r}_{\text{patch}}^t \right], \tag{8}$$

where r_{cls}^t denotes the global frame token, r_{reg}^t denotes the retained register tokens, and $\bar{r}_{\text{patch}}^t$ is obtained by averaging all patch tokens of the t-th frame. Stacking the compact token sets from all sampled frames gives the visual foundation embedding set:

$$R_v = \mathcal{E}_v(I_v) = \{R_v^t\}_{t=1}^{T_v}. \tag{9}$$

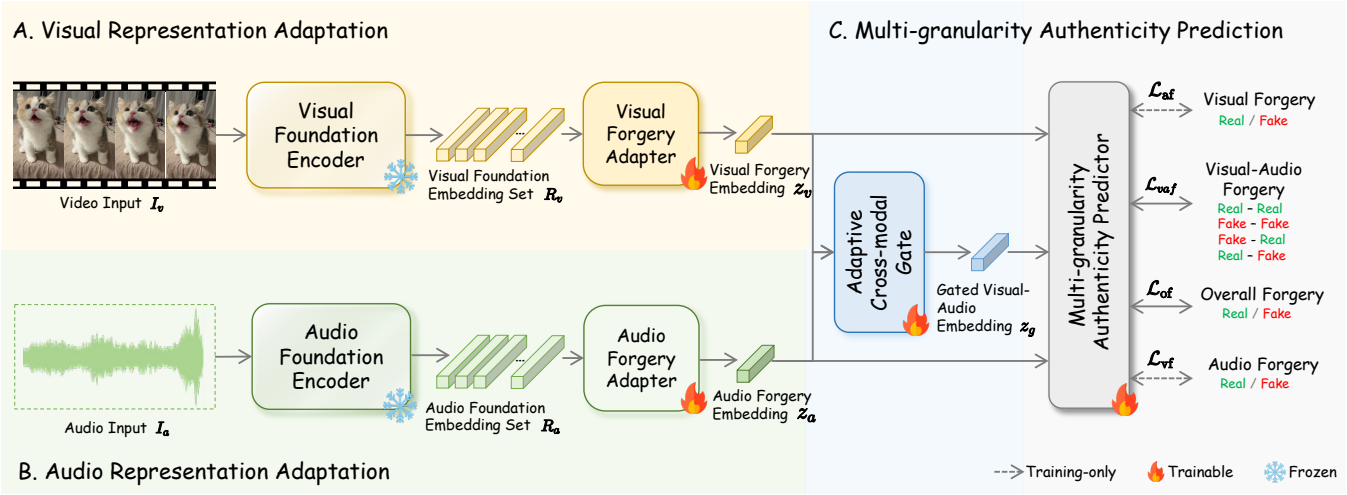

Figure 1: Overview of GEF-VA. Visual and audio foundation embedding sets are adapted into modality-specific forgery embeddings. An adaptive cross-modal gate produces a gated video-audio embedding, and a multi-granularity predictor learns modality-level, video-audio, and overall forgery objectives.

This construction preserves complementary visual evidence across frames, covering both global video semantics and localized appearance variations.

We then summarize R_v into a video-level evidence vector with distributional pooling. Concretely, we flatten all frame-token embeddings and concatenate four statistics: mean, standard deviation, maximum, and minimum:

$$h_v = [mean(R_v); std(R_v); max(R_v); min(R_v)]\,. \quad (10)$$

The mean captures the average visual response, the standard deviation reflects variation across frames and tokens, while the maximum and minimum preserve extreme activation patterns that may correspond to localized or rare generation artifacts. Compared with a single pooled feature, this statistical summary retains richer evidence distribution from the visual stream.

Finally, the summarized visual evidence is transformed by a lightweight Visual Forgery Adapter:

$$z_v = A_v(h_v), \quad (11)$$

where A_v denotes the trainable adapter and z_v is the visual forgery representation. This design keeps the foundation encoder frozen while converting general-purpose visual representations into forgery-sensitive visual evidence.

4.2 Audio Representation Adaptation

Audio evidence in generated media is often distributed across multiple levels, from global acoustic semantics to local temporal variation and spectral irregularities. To preserve such information, we first decode and normalize the audio input I_a, convert it into a time-frequency representation, and feed it into a frozen audio foundation encoder $\mathcal{E}_a$.

For each audio sample, we retain both the global audio representation and patch-level representations:

$$R_a = [r_{\text{pool}}; r_{\text{patch}}]\,, \quad (12)$$

where r_{pool} denotes the global audio embedding and r_{patch} denotes the retained patch-level audio embeddings. The global representation captures overall acoustic characteristics, while patch-level embeddings preserve localized temporal and spectral responses.

We then summarize R_a into an audio-level evidence vector with distributional pooling. Concretely, we concatenate the global representation with four statistics of the patch-level embeddings: mean, standard deviation, maximum, and minimum:

$$h_a = \big[r_{\text{pool}}; \text{mean}(r_{\text{patch}}); \text{std}(r_{\text{patch}}); \\ \max(r_{\text{patch}}); \min(r_{\text{patch}})\big]. \quad (13)$$

Finally, the summarized audio evidence is transformed by a lightweight Audio Forgery Adapter:

$$z_a = A_a(h_a), \quad (14)$$

where A_a denotes the trainable adapter and z_a is the audio forgery representation. This design keeps the foundation encoder frozen while converting general-purpose audio representations into forgery-sensitive acoustic evidence.

4.3 Multi-granularity Authenticity Prediction

After obtaining the visual forgery representation z_v and the audio forgery representation z_a, the model needs to combine modality-specific evidence with cross-modal relations. Since different samples may rely on different evidence sources, we introduce an adaptive cross-modal gate to estimate the relative contribution of visual and audio evidence for each input:

$$g = \sigma\left(\mathcal{G}\left(z_v, z_a\right)\right), \quad (15)$$

where $\mathcal{G}$ denotes the gating model and $\sigma(\cdot)$ is the sigmoid function. The gated video-audio representation is computed as

$$z_g = g \odot z_v + (1 - g) \odot z_a, \quad (16)$$

where $\odot$ denotes element-wise multiplication. This gate allows the model to adjust the contribution of each modality according to the input sample, rather than relying on a fixed fusion rule.

To better capture the relationship between visual and audio evidence, we further construct a fusion input that combines modality evidence, gated evidence, and explicit cross-modal comparison terms:

$$z = [z_v; z_a; z_g; z_v \odot z_a; |z_v - z_a|; z_v - z_a]. \tag{17}$$

Here, $z_v \odot z_a$ reflects evidence co-activation, $|z_v - z_a|$ measures the magnitude of cross-modal discrepancy, and $z_v - z_a$ preserves the direction of the discrepancy.

Based on the relational evidence representation z, we perform video-audio-level authenticity prediction with lightweight classification heads. Specifically, the four-class configuration head predicts the video-audio authenticity type, and the binary head predicts the overall authenticity:

$$\ell_c = \mathcal{H}_c(z) \in \mathbb{R}^4, \qquad \ell_b = \mathcal{H}_b(z) \in \mathbb{R}^2, \tag{18}$$

where ℓ_c denotes the four-class logits and ℓ_b denotes the binary logits. In addition, we attach two modality-level auxiliary heads to the original visual and audio representations:

$$\ell_v = \mathcal{H}_v(z_v) \in \mathbb{R}^2, \qquad \ell_a = \mathcal{H}_a(z_a) \in \mathbb{R}^2. \tag{19}$$

Here, ℓ_v and ℓ_a denote the modality-level binary logits, and $\mathcal{H}_v$ and $\mathcal{H}_a$ are used only during training to encourage each modality branch to retain discriminative forgery evidence before cross-modal prediction.

$$\mathcal{L} = \mathcal{L}_{vaf} + \lambda_b \mathcal{L}_{of} + \lambda_m(\mathcal{L}_{vf} + \mathcal{L}_{af}). \tag{20}$$

All terms are implemented with cross-entropy losses, where the four-class configuration loss $\mathcal{L}_{vaf}$ supports video-audio authenticity prediction, the binary loss $\mathcal{L}_{of}$ supports overall authenticity prediction, and the auxiliary modality losses $\mathcal{L}_{vf}$ and $\mathcal{L}_{af}$ keep the visual and audio representations discriminative before fusion.

4.4 Training Strategy

To improve GEF-VA's robustness to class imbalance and ambiguous video-audio authenticity configurations, we adopt a two-stage training strategy. In the first stage, all trainable components are optimized on the full training set using the multi-granularity loss. Exposure to the complete training distribution allows the visual adapter, audio adapter, cross-modal gate, and prediction heads to first learn general modality-specific and cross-modal evidence patterns.

In the second stage, we refine the model using hard samples identified by the first-stage predictor. Specifically, we evaluate all training samples with the model and regard samples receiving lower confidence for their ground-truth configuration labels as more difficult. To prevent the refinement subset from being dominated by classes, we rank the samples separately within each of the four configuration classes and select the 1000 most difficult samples from each class, forming a class-balanced hard-sample subset.

The model is then fine-tuned on this subset using the same multi-granularity loss. This class-balanced refinement concentrates optimization on ambiguous and configuration-sensitive cases while ensuring that all four authenticity configurations are sufficiently represented. It therefore strengthens the decision boundaries between easily confused configurations while retaining modality-level and overall authenticity supervision. During inference, the refined model is directly used for the final binary and four-class predictions.

5 Experiments

5.1 Experimental Setup

Dataset and Evaluation Protocol. We conduct experiments on MVAD [Hu *et al.*, 2025], a general-purpose multimodal video-audio dataset for AI-generated content detection. MVAD contains four video-audio authenticity configurations: fake video–fake audio (FF), fake video–real audio (FR), real video–fake audio (RF), and real video–real audio (RR). It covers diverse visual styles and content categories, making it suitable for evaluating multimodal detectors across diverse video-audio scenarios. MVAD provides an official training set and an official test set. We randomly split the official training set into training and internal validation subsets at a ratio of 8:2. Unless otherwise specified, all experiments are conducted on the internal validation subset. The official test set exhibits a substantially different data distribution and is reserved for final challenge evaluation. To characterize the model's performance under these distinct evaluation conditions, we report the internal validation and official test results separately.

Evaluation Metrics. We evaluate our method on both binary classification and four-class video-audio authenticity classification. For each task, we report AUC, ACC, AP, and F1. The task score is computed by assigning 20% weight to AUC, 30% to ACC, 30% to AP, and 20% to F1. The final score combines the binary classification score and the four-class classification score, with weights of 40% and 60%, respectively. For single-modality detectors, we only evaluate single-modality binary classification, where each model predicts whether its corresponding modality is real or fake rather than distinguishing joint video-audio authenticity configurations.

Baseline Methods. To provide a comprehensive evaluation, we compare GEF-VA with three categories of detection paradigms. The first category includes AI-generated video detectors, such as DeMamba [Chen *et al.*, 2024] and NSG-VD [Zhang *et al.*, 2025b], which rely only on visual evidence. The second category includes AI-generated audio detectors, including Tran et al. [Tran *et al.*, 2025] and WPT-SSL [Xie *et al.*, 2026], which focus on acoustic forgery cues. The third category includes video-audio detectors, such as AV-Graphs [Yin *et al.*, 2024] and AVFF [Oorloff *et al.*, 2024], which exploit multimodal information for authenticity prediction. For a fair comparison, all methods are trained on the same training set.

Implementation Details. The visual foundation encoder $\mathcal{E}_v$ is instantiated with DINOv3 ViT-7B/16 [Siméoni *et al.*, 2025], while the audio foundation encoder $\mathcal{E}_a$ is instantiated with Dasheng-1.2B [Dinkel *et al.*, 2024]. Both encoders remain frozen throughout training. For each video, we uniformly sample 16 frames and retain compact visual

Method	Binary Classification					Four-Class Classification					Final Score
	AUC	ACC	AP	F1	Score	AUC	ACC	AP	F1	Score	
Internal Validation Subset											
AI-Generated Video Detectors											
DeMamba	96.79	93.32	97.76	92.07	95.10	–	–	–	–	–	–
NSG-VD	99.84	98.79	99.20	95.61	98.49	–	–	–	–	–	–
AI-Generated Audio Detectors											
Tran et al.	91.39	86.72	78.13	76.64	83.06	–	–	–	–	–	–
WPT-SSL	96.21	93.36	91.31	85.71	91.79	–	–	–	–	–	–
AI-Generated Video-Audio Detectors											
AVGraphs	99.77	97.64	99.50	96.16	98.32	99.43	95.11	96.71	92.17	95.87	96.85
AVFF	98.60	97.72	98.42	97.55	98.07	98.10	96.70	98.02	96.40	97.32	<u>97.62</u>
Ours	100.0	99.85	100.0	99.89	99.93	99.99	99.66	99.98	99.66	99.82	**99.87**
Official DDL 2.0 Test Set											
Ours*	92.88	91.06	95.91	91.06	92.88	85.45	88.07	70.96	68.56	78.51	84.26

Table 1: Overall results under the internal validation and official DDL 2.0 test protocols. Video and audio detectors are evaluated on their corresponding modality-level binary authenticity tasks, whereas video-audio detectors are evaluated on both overall binary authenticity prediction and four-class configuration classification. For each applicable task, we report AUC (%), ACC (%), AP (%), F1 (%), and the corresponding Score. Ours* reports the result of our challenge submission on the official test set; all other results are evaluated on the internal validation subset. The best and second-best validation results are marked in **bold** and <u>underline</u>, respectively.

Module	Score
DINOv3 ViT-L/16	99.07
XLS-R	99.51
Late fusion	93.86
Concat. fusion	96.77
Average fusion	95.21
Gated fusion	96.94
Single-stage	98.71
Ours	**99.87**

(a) Modality evidence integration. Adaptive fusion and hard-sample refinement enable GEF-VA to better exploit complementary visual and audio foundation cues under diverse forgery configurations.

Module	Score
w/o V-Adapter	97.24
w/o A-Adapter	98.46
w/o Gate	96.40
Ours	**99.87**

(b) Core module ablations.

V-Set	A-Set	Stat.	Score
✗	✗	✗	96.32
✓	✗	✓	98.10
✗	✓	✓	98.44
✓	✓	✗	97.25
✓	✓	✓	**99.87**

(c) Representation design.

Table 2: Ablation study of GEF-VA. We evaluate the contributions of key design choices and report the Final Score.

tokens consisting of the CLS token, register tokens, and a mean-pooled patch token. For audio, we decode each audio track as a mono waveform at 16 kHz, crop or repeat it to a fixed duration, normalize it, and convert it into a log-mel spectrogram. Both pooled and patch-level audio representations are retained. The visual and audio adapters are implemented as two-layer MLPs with GELU activation, transforming the summarized foundation representations into 256-dimensional modality-specific forgery representations. The adaptive cross-modal gate concatenates the visual and audio forgery representations and processes them with a lightweight MLP followed by a sigmoid function to generate a 256-dimensional weight vector for feature-wise modality fusion. The trade-off coefficients are set to $\lambda_b = 0.35$ and $\lambda_m = 0.15$. We first train the model on the full training set and then fine-tune it on a class-balanced hard-sample subset. Based on the prediction confidence from the first stage, we select 1000 difficult samples from each of the four configuration classes. AdamW is used in both stages. The model is trained for 10 epochs with a learning rate of 1×10^{-4} in the first stage and fine-tuned for 1 epoch at 3×10^{-5} in the second stage. The batch size is 64, and all training is conducted on one NVIDIA A800 GPU.

5.2 Comparison with the State of the Art

Table 1 compares GEF-VA with three categories of methods, including AI-generated video detectors, AI-generated audio detectors, and AI-generated video-audio detectors. The video and audio detectors are evaluated on their corresponding modality-level binary authenticity tasks, while the video-audio detectors are additionally evaluated on four-class video-audio configuration classification. All methods are trained on the same training set, and the results are evaluated on the internal validation subset unless otherwise specified. Among the single-modality baselines, NSG-VD achieves the highest Binary Score of 98.49%, demonstrating the strong discriminative capability of visual evidence for video authenticity prediction. GEF-VA achieves a Binary Score of 99.93% on overall video-audio authenticity prediction. For four-class configuration classification, GEF-VA obtains a score of 99.82%, exceeding the strongest video-audio baseline, AVFF, by 2.50 percentage points. GEF-VA further achieves the highest Final Score of 99.87%, representing an improvement of 2.25 percentage points over AVFF. These results demonstrate the effectiveness of preserving modality-specific evidence and performing configuration-aware gated fusion for fine-grained video-audio authenticity prediction. We further report the performance of our DDL 2.0 Challenge submission on the official test set. GEF-VA obtains a Binary Score of 92.88%, a Four-Class Score of 78.51%, and a Final Score of

84.26%. The lower performance relative to the internal validation results indicates a notable gap between the two evaluation settings, which may be related to the distribution difference between the internal validation subset and the official test set.

5.3 Ablation Study

We conduct ablation studies on the validation set and report the Final Score in Table 2.

Analysis of Backbone, Fusion, and Training Strategy. Table 2(a) analyzes the effects of backbone selection, fusion strategy, and training scheme. Replacing the foundation encoder with DINOv3 ViT-L/16 or XLS-R only causes moderate performance drops, indicating that the effectiveness of GEF-VA does not simply come from relying on a specific powerful foundation encoder. Instead, the proposed evidence adaptation and fusion design can consistently exploit different visual and audio foundation representations. In contrast, generic fusion strategies, including late fusion, concatenation, averaging, and vanilla gated fusion, lead to more evident degradation, suggesting that naive multimodal aggregation may introduce modality interference rather than fully exploiting complementary cues. GEF-VA achieves the best Final Score of 99.87, demonstrating the effectiveness of configuration-aware evidence fusion. Compared with the single-stage variant, the full two-stage training strategy further improves performance, showing that hard-sample refinement helps enhance discrimination under ambiguous authenticity configurations.

Effectiveness of Core Modules. Removing either the Visual Forgery Adapter or the Audio Forgery Adapter degrades performance, indicating that modality-specific adaptation is necessary for converting frozen foundation features into forgery-sensitive evidence. The largest drop occurs when the adaptive gate is removed, showing that sample-dependent evidence allocation is crucial for video-audio forgery detection, where decisive cues may come from different modalities under different configurations.

Analysis of Representation Design. The representation ablations validate the importance of preserving foundation embedding sets and their distributional statistics. Removing the visual set, audio set, or statistical pooling consistently reduces the Final Score, while the full design achieves the best performance. These results show that GEF-VA benefits not only from multimodal fusion, but also from retaining rich evidence distributions before fusion, which helps capture both global modality cues and localized forgery traces.

6 Conclusion

We present **GEF-VA**, a configuration-aware evidence fusion framework for general AI-generated video-audio detection. By integrating modality-preserving representation adaptation, gated cross-modal evidence fusion, and multi-granularity authenticity prediction, GEF-VA jointly learns modality-level, video-audio configuration-level, and overall authenticity objectives through multi-granularity supervision. Experiments on the DDL 2.0 challenge validate its effectiveness for both binary authenticity prediction and four-class video-audio configuration recognition. We hope this work offers new insights into leveraging pretrained foundation models for configuration-aware multimodal forensics.

Acknowledgments

This work is supported by the Guangdong Science and Technology Department (No. 2024ZDZX2004) and the Program of Guangdong Education Department.

Contribution Statement

†: Shuaibo Li and Laixin Zhang contributed equally.
∗: Lei Zhu is the corresponding author.

References

[Agarwal *et al.*, 2020] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020.

[Babu *et al.*, 2022] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*, 2022.

[Bai *et al.*, 2024] Jianfa Bai, Man Lin, and Gang Cao. Ai-generated video detection via spatio-temporal anomaly learning. *arXiv preprint arXiv:2403.16638*, 2024.

[Borsos *et al.*, 2023] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

[Chang *et al.*, 2024] Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. What matters in detecting ai-generated videos like sora? *arXiv e-prints*, pages arXiv–2406, 2024.

[Chen *et al.*, 2024] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024.

[Chugh *et al.*, 2020] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 439–447, 2020.

[Dinkel *et al.*, 2024] Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up masked audio encoder learning for general audio classification. *arXiv preprint arXiv:2406.06992*, 2024.

[Gu and Dao, 2024] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conf. on Lang. Modeling*, 2024.

[Gu *et al.*, 2021] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021.

[Hu *et al.*, 2025] Mengxue Hu, Yunfeng Diao, Changtao Miao, Jianshu Li, Zhe Li, and Joey Tianyi Zhou. Mvad: A comprehensive multimodal video-audio dataset for aigc detection. *arXiv preprint arXiv:2512.00336*, 2025.

[Ilyas *et al.*, 2023] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, 136:110124, 2023.

[Khalid *et al.*, 2021a] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, pages 7–15, 2021.

[Khalid *et al.*, 2021b] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[Kundu *et al.*, 2025] Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28050–28060, 2025.

[Li *et al.*, 2018] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018.

[Li *et al.*, 2021] Shuaibo Li, Shibiao Xu, Wei Ma, and Qiu Zong. Image manipulation localization using attentional cross-domain cnn features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5614–5628, 2021.

[Li *et al.*, 2024] Shuaibo Li, Wei Ma, Jianwei Guo, Shibiao Xu, Benchong Li, and Xiaopeng Zhang. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12523–12533, 2024.

[Li *et al.*, 2025] Shuaibo Li, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, and Lei Zhu. Toward medical deepfake detection: A comprehensive dataset and novel method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–637. Springer, 2025.

[Li *et al.*, 2026a] Shuaibo Li, Yijun Yang, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, Qing Zhang, and Lei Zhu. Synerdetect: Hierarchical synergistic learning for generalizable ai-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6405–6414, 2026.

[Li *et al.*, 2026b] Shuaibo Li, Laixin Zhang, Wei Ma, Jianwei Guo, Shibiao Xu, Zhijie Qiu, and Hongbin Zha. Realnet: Efficient and unsupervised detection of ai-generated images via real-only representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38889–38898, 2026.

[Ma *et al.*, 2025] Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting ai-generated video via frame consistency. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[Martín-Doñas and Álvarez, 2022] Juan M. Martín-Doñas and Aitor Álvarez. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9241–9245, 2022.

[Miao *et al.*, 2025] Hui Miao, Yuanfang Guo, Zeming Liu, and Yunhong Wang. Multi-modal deepfake detection via multi-task audio-visual prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 612–621, 2025.

[Mittal *et al.*, 2020] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020.

[Oorloff *et al.*, 2024] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024.

[Peng *et al.*, 2024] Chunlei Peng, Zimin Miao, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Where deepfakes gaze at? spatial-temporal gaze inconsistency analysis for video face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:4507–4517, 2024.

[Pratap *et al.*, 2024] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97), 2024.

[Schneider *et al.*, 2019] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[Siméoni *et al.*, 2025] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo

Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[Song *et al.*, 2024] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. *Advances in Neural Information Processing Systems*, 37:122054–122077, 2024.

[Tran *et al.*, 2025] Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, and David Guennec. Multi-level ssl feature gating for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11766–11775, 2025.

[Xie *et al.*, 2026] Yuankun Xie, Ruibo Fu, Xiaopeng Wang, Zhiyong Wang, Songjun Cao, Long Ma, Haonan Cheng, and Long Ye. Detect all-type deepfake audio: Wavelet prompt tuning for enhanced auditory perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35922–35930, 2026.

[Xu *et al.*, 2023] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023.

[Yang *et al.*, 2023] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.

[Yin *et al.*, 2024] Qilin Yin, Wei Lu, Xiaochun Cao, Xiangyang Luo, Yicong Zhou, and Jiwu Huang. Fine-grained multimodal deepfake classification via heterogeneous graphs. *International Journal of Computer Vision*, 132(11):5255–5269, 2024.

[Zhang *et al.*, 2024] Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised XLS-r and SLS classifier. In *ACM Multimedia 2024*, 2024.

[Zhang *et al.*, 2025a] Laixin Zhang, Shuaibo Li, Wei Ma, and Hongbin Zha. Truemoe: Dual-routing mixture of discriminative experts for synthetic image detection. *arXiv preprint arXiv:2509.15741*, 2025.

[Zhang *et al.*, 2025b] Shuhai Zhang, Zihao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for ai-generated video detection. In *Advances in Neural Information Processing Systems*, 2025.

[Zheng *et al.*, 2025] Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free ai-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[Zou *et al.*, 2024] Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan. Cross-modality and within-modality regularization for audio-visual deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4900–4904. IEEE, 2024.