Sparse Visual Expert Activation and Structured Fusion for AI-Generated Audio-Video Detection

Haoyu Wang[1] , **Kui Ren**[1,2] , **Zhongjie Ba**[1,2] , **Haohao Li**[1] , **Junhao Wang**[1] and **Hongshuo Jin**[1] , **Peng Cheng**[1,2] , **Li Lu**[1,2] , **Qing Wen**[1] , **Zhan Qin**[1]

[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University

[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

{haoyuwang2002, kuiren, zhongjieba, wangjunhao, peng_cheng, li.lu, qinzhan}@zju.edu.cn, lih32768@gmail.com, 994195842@qq.com, qwerwenging@icloud.com

Abstract

AI-generated audio-video synthesis can produce media in which the visual stream, the audio stream, or both are manipulated. We formulate AI-generated audio-video detection as structured prediction over visual and audio authenticity variables, preserving the semantics of the four joint states: real–real, fake–real, real–fake, and fake–fake. We present a single integrated detector comprising a compact fixed visual expert ensemble, an XLS-R–LoRA–Mamba audio branch, and a learned log-linear posterior fusion head. The two branches estimate modality-wise authenticity posteriors, which are combined using global modality weights and state-specific offsets for four-state prediction. All retained visual experts are evaluated for each input; the method does not use dynamic expert routing. Deterministic audio-video input preparation is applied before modality-specific inference. Under the official Track 2 evaluation, the final submitted model achieves a binary score of 0.9396, a four-class score of 0.7746, and a weighted score of 0.8406. We additionally provide a descriptive media-quality audit of the released evaluation package; this analysis does not constitute a direct robustness evaluation.

1 Introduction

Recent advances in AI-generated audiovisual synthesis have made it increasingly easy to create realistic videos, speech, and synchronized multimedia content. While these capabilities support creative and productive applications, they also complicate the assessment of media authenticity. In particular, manipulated audiovisual content need not alter both modalities simultaneously: a video may contain manipulated visual content with authentic audio, authentic visual content with manipulated audio, or manipulation in both streams. Detecting that a sample is manipulated is therefore not sufficient when the practical goal is also to identify which modality has been affected.

This setting can be represented by two modality-wise authenticity variables: visual authenticity and audio authenticity. Their joint states correspond to real visual–real audio (RR), fake visual–real audio (FR), real visual–fake audio (RF), and fake visual–fake audio (FF). The FR and RF cases are particularly important because they require the detector to distinguish modality-specific manipulation from joint manipulation. Treating these four states as unrelated classes can obscure their underlying structure. In contrast, modeling them through visual and audio authenticity variables preserves the semantic relation between the joint label and its two constituent modalities. This formulation is consistent with the general audio-video detection setting considered in MVAD and the IJCAI 2026 Track 2 benchmark [HuMengXue0104, 2026; DDL 2.0 Organizing Committee, 2026].

Existing approaches to AI-generated media detection often focus on a single modality, such as visual generation traces in images or videos, or acoustic artifacts in synthesized speech. Such detectors can be effective when manipulation is confined to their respective input stream, but they do not directly determine whether a suspicious audio-video sample is visually manipulated, acoustically manipulated, or manipulated in both modalities. Multimodal approaches can combine information from the two streams, yet the fusion mechanism may obscure how each modality contributes to the final decision. This limitation is especially relevant for partial manipulation, where a strong signal in one modality should not automatically imply that the other modality is also manipulated.

We address this problem through structured posterior prediction. Instead of directly learning four independent class scores from a joint audio-video representation, SPARSEAVF first estimates a visual authenticity posterior and an audio authenticity posterior. The visual branch uses a compact fixed ensemble of visual experts, while the audio branch combines XLS-R representations, LoRA adaptation, and Mamba-based temporal modeling. Their outputs are subsequently combined by a learned log-linear posterior fusion head with global modality weights and state-specific offsets. The fusion design is intentionally low-capacity: it does not use sample-dependent routing, adaptive modality reliability estimation, audio-video synchronization features, or correspondence modeling. Rather, it provides a structured mechanism for mapping modality-wise authenticity evidence to the four joint states.

The visual expert ensemble is fixed and all retained experts

are evaluated for each input. Thus, the system does not rely on input-conditioned sparse routing or Top-K expert selection. We use the term "compact" to describe the bounded, preselected expert set and its role in reducing reliance on a single visual decision boundary. Before modality-specific inference, the system applies deterministic audio-video input preparation, including standardized visual frame processing and waveform preparation. These operations are intended to produce consistent inputs for the visual and audio branches; they are not presented as a direct robustness guarantee.

The final system is submitted as one integrated audio-video detector. Its official Track 2 evaluation yields a binary score of 0.9396, a four-class score of 0.7746, and a weighted score of 0.8406. In addition to reporting the final detection result, we provide a descriptive media-quality audit of the released evaluation package. The audit characterizes variation in visual and acoustic quality indicators, but it is not used as a training signal, a model-selection criterion, or a direct robustness evaluation.

Our contributions are summarized as follows:

- We formulate AI-generated audio-video detection as structured prediction over visual and audio authenticity variables, preserving the semantics of the RR, FR, RF, and FF states.
- We develop a single integrated detector that combines a compact fixed visual expert ensemble, an XLS-R–LoRA–Mamba audio branch, and learned log-linear posterior fusion.
- We distinguish deterministic media preparation, modality-specific authenticity prediction, and structured posterior fusion, and report a descriptive audit of media-quality variation in the released evaluation package.

2 Related Work

2.1 Unimodal AI-Generated Content Detection

AI-Generated Image Detection

AI-generated image detection has examined a range of generator-induced signals, including model-specific artifacts, frequency-domain traces, and representation-level cues. Early work showed that discriminative detectors can identify synthetic-image traces beyond the generators observed during training when suitable data variation is available [Wang *et al.*, 2020]. Frequency-based studies further investigated spectral deviations associated with common image-generation pipelines [Frank *et al.*, 2020; Durall *et al.*, 2020]. More recent work has focused on transfer to previously unseen generators. For example, Ojha et al. leverage pretrained visual representations for improved cross-generator generalization [Ojha *et al.*, 2023], while DIRE uses the discrepancy between an input image and its diffusion-based reconstruction as a detection signal [Wang *et al.*, 2023].

These studies indicate that visual authenticity evidence can vary with the generator family, content domain, representation, and post-processing conditions. Our visual branch therefore uses a compact fixed ensemble of visual experts to aggregate complementary visual predictions. All retained experts are evaluated for every input; the design is a fixed ensemble rather than an input-conditioned routing mechanism. Visual-only approaches, however, cannot determine whether the audio stream of an audio-video sample is authentic.

AI-Generated Audio Detection

AI-generated audio detection has been extensively studied through spoofing benchmarks such as ASVspoof, which include variation in synthesis, conversion, replay, recording, and transmission conditions [Yamagishi *et al.*, 2021]. Many systems combine spectral and temporal cues to distinguish bona fide speech from manipulated or synthesized audio. AASIST, for instance, models heterogeneous spectral and temporal artifacts through graph-attention-based aggregation [Jung *et al.*, 2021].

Self-supervised speech models have also become useful transfer backbones for audio authenticity prediction. XLS-R provides cross-lingual contextual speech representations learned from large-scale self-supervised pretraining [Babu *et al.*, 2021], while LoRA enables parameter-efficient adaptation of pretrained linear transformations [Hu *et al.*, 2021]. Selective state-space models such as Mamba offer an alternative mechanism for sequence aggregation over long acoustic representations [Gu and Dao, 2023]. In our system, these components are used to construct an audio authenticity branch. Audio-only prediction remains insufficient for the present task because it cannot attribute manipulation to the visual stream.

2.2 Multimodal AI-Generated Audio-Video Detection

Multimodal deepfake benchmarks have increasingly considered settings in which visual and audio streams may be manipulated jointly or independently. FakeAVCeleb introduced paired audio-video deepfake data and evaluated unimodal, ensemble-based, and multimodal detectors [Khalid *et al.*, 2021b; Khalid *et al.*, 2021a]. AV-Deepfake1M further includes audio-only, visual-only, and jointly manipulated examples at larger scale [Cai *et al.*, 2023]. The MVAD benchmark and its released competition resources extend this setting to general AI-generated audio-video content across realistic and anime styles, multiple content categories, and four modality-composition states [HuMengXue0104, 2026].

A common direction in multimodal detection is to learn correspondence or feature interactions between audio and visual streams. AVFF, for example, learns audio-visual feature representations before supervised deepfake classification [Oorloff *et al.*, 2024]. Such approaches can exploit cross-stream inconsistency when audiovisual correspondence is informative. Our formulation addresses a related but more restricted question: whether visual authenticity, audio authenticity, or both should be assigned as manipulated. It first estimates modality-wise authenticity posteriors and subsequently maps them to the joint RR, FR, RF, and FF states using a low-capacity log-linear posterior fusion head.

Accordingly, our method does not claim to model sample-specific audio-video synchronization, correspondence, or modality reliability. Instead, it preserves the semantic structure of partially manipulated samples by making the visual

and audio authenticity variables explicit before joint-state prediction.

3 Method

3.1 Problem Formulation

Let $x = (v, a)$ denote an audio-video sample, where v and a denote the visual and audio streams, respectively. We associate each sample with a visual authenticity variable $y_v \in \{R, F\}$ and an audio authenticity variable $y_a \in \{R, F\}$. The joint authenticity label is defined as

$$y = (y_v, y_a) \in \{\mathrm{RR}, \mathrm{FR}, \mathrm{RF}, \mathrm{FF}\}, \tag{1}$$

where the first symbol denotes visual authenticity and the second symbol denotes audio authenticity. Thus, FR denotes manipulated visual content paired with authentic audio, whereas RF denotes authentic visual content paired with manipulated audio. The binary target is derived from the joint state:

$$y_{\mathrm{bin}} = \mathbb{I}[y \neq \mathrm{RR}]. \tag{2}$$

Rather than treating the four joint states as unrelated classes, SPARSEAVF first estimates modality-wise authenticity posteriors $p_v(y_v \mid v)$ and $p_a(y_a \mid a)$, and subsequently predicts the joint posterior $p(y_v, y_a \mid v, a)$. This formulation preserves the semantic relation between partial manipulation and its affected modality, while allowing visual, audio, and fusion errors to be inspected separately.

3.2 Framework Overview

Figure 1 presents the overall architecture. Raw audio-video input first undergoes robustness-oriented audio–visual preparation. The visual branch estimates visual authenticity from prepared video frames using a compact fixed set of visual experts, whereas the audio branch estimates audio authenticity from a prepared waveform. A low-capacity posterior fusion head subsequently combines the modality-wise posteriors and predicts the four joint authenticity states.

The preparation stage separates deterministic media normalization, decoder-failure handling, and stochastic audio augmentation used during training only. The modality-specific branches are optimized with binary authenticity labels. Their posterior estimates are subsequently used by the fusion head, which is optimized with four-state labels.

Robustness-Oriented Audio–Visual Preparation. AI-generated traces may be affected by heterogeneous decoding, spatial resolution, temporal duration, channel response, and media-delivery conditions. We therefore apply an input preparation procedure before modality-specific inference. The procedure standardizes media representations and exposes the audio branch to selected training-time perturbations. It does not alter the four-state label definition or provide a direct guarantee of robustness under arbitrary real-world degradation.

Visual-stream preparation. Let V denote a decoded video containing N frames. For a target length T, the visual pipeline selects uniformly distributed frame indices:

$$\tau_t = \left\lfloor \frac{t-1}{\max(T-1, 1)} (N-1) \right\rfloor, \qquad t = 1, \ldots, T. \tag{3}$$

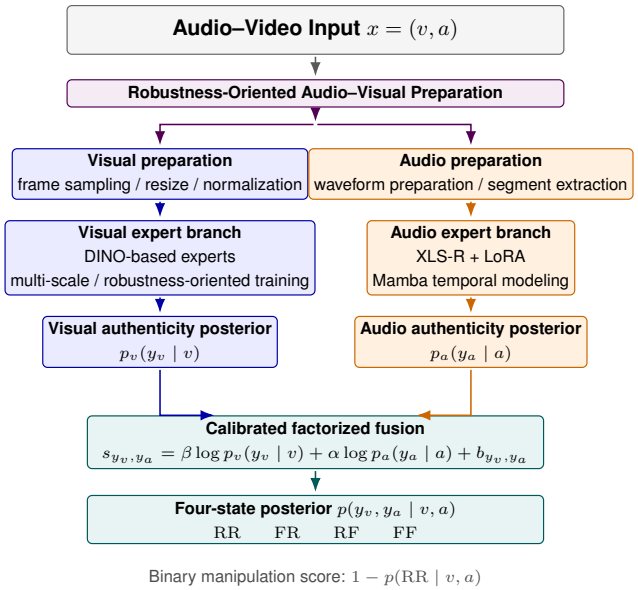

Figure 1: Overview of SPARSEAVF. Robustness-oriented audio–visual preparation standardizes video and waveform inputs before modality-specific inference. The visual and audio branches estimate authenticity posteriors independently, and posterior fusion maps them to the four joint states.

For each selected index, the implementation converts the decoded BGR frame to RGB, applies linear-interpolation resizing, and normalizes pixel values to $[0, 1]$:

$$\tilde{v}_t = \frac{1}{255} \mathrm{Resize}_{H \times W}^{\mathrm{linear}} \left(\mathrm{RGB} \left(\mathrm{Decode}(V, \tau_t) \right) \right). \tag{4}$$

The final visual pipeline uses $H = W = 384$. If an individual frame cannot be decoded, the most recent valid frame is reused when available. If no valid frame is available, a zero-valued frame is inserted to preserve a well-defined input tensor. These fallback rules are deterministic input handling operations rather than data augmentation.

The resulting visual input is

$$\tilde{v} = [\tilde{v}_1, \ldots, \tilde{v}_T] \in \mathbb{R}^{T \times 3 \times H \times W}. \tag{5}$$

This procedure standardizes frame count, spatial resolution, channel ordering, and invalid-frame handling before visual authenticity prediction.

Audio-stream preparation and augmentation. The audio pipeline first decodes each signal as a mono 16-kHz waveform. During audio-expert training, stochastic waveform augmentation is applied before fixed-length segment construction. The augmentation policy samples at most one operation from each of three ordered groups:

$$\hat{a} = \mathcal{A}_C(\mathcal{A}_B(\mathcal{A}_A\left(\mathrm{Mono}_{16k}(a)\right))), \tag{6}$$

where $\mathcal{A}_A$ denotes channel and filtering perturbations, $\mathcal{A}_B$ denotes additive-noise perturbations, and $\mathcal{A}_C$ denotes mask-based perturbations.

The first group samples from RawBoost-style convolutive or impulsive distortion, low-pass filtering, band-pass filtering,

and room impulse response convolution. The second group samples from white Gaussian noise, low-frequency rumble noise, and background-noise mixing. The third group applies either no operation or a SpecAugment-style masking operation. Restricting the policy to at most one perturbation from each group limits the number of simultaneously applied transformations while exposing the audio branch to variation in channel response, noise floor, and local signal observability. The original authenticity label is retained when an augmentation is sampled.

After augmentation, the waveform is converted to a fixed-duration segment. Let $L_a = 4 \times 16000$ denote the target number of samples. For a waveform shorter than L_a, repetition and truncation are used:

$$\tilde{a} = \mathrm{RepeatTrim}(\hat{a}, L_a). \tag{7}$$

For a waveform longer than L_a, a contiguous segment is selected:

$$\tilde{a} = \mathrm{Crop}(\hat{a}, s, L_a), \tag{8}$$

where s is sampled from the valid range of segment starts during training. At evaluation time, stochastic augmentation is disabled, while mono conversion, resampling, and fixed-duration waveform construction are retained.

3.3 Visual Expert Model

Visual traces associated with AI-generated images and videos can vary across generation pipelines, content types, and post-processing conditions. To reduce reliance on a single visual decision boundary, the visual branch contains a compact fixed set of DINO-based expert classifiers:

$$\mathcal{S} = \{E_1, \ldots, E_M\}. \tag{9}$$

The retained experts differ in their pre-specified model scale and training configuration. All M experts are evaluated for every input. Thus, the ensemble does not use input-conditioned routing, Top-K selection, or dynamic expert activation.

For a prepared frame $\tilde{v}_t$, expert E_m produces a manipulated-image probability:

$$q_{t,m} = \sigma\big(g_m(E_m(\tilde{v}_t))\big), \tag{10}$$

where g_m is a binary classification head. Frame-level scores are pooled within each expert:

$$q_m = \mathcal{P}(\{q_{t,m}\}_{t=1}^{T}), \tag{11}$$

where $\mathcal{P}$ denotes the configured temporal pooling rule. The expert outputs are then combined through a fixed aggregation function $\mathcal{A}$:

$$\begin{aligned} q_v &= \mathcal{A}(q_1, \ldots, q_M), \\ p_v(F \mid v) = q_v, &\qquad p_v(R \mid v) = 1 - q_v. \end{aligned} \tag{12}$$

Visual expert training. Let $\mathcal{D}_v = \{(V_i, y_v^{(i)})\}_{i=1}^{N_v}$ denote the visual training set, where $\tilde{y}_v^{(i)} = \mathbb{I}[y_v^{(i)} = F]$ is the corresponding binary visual-authenticity label. Each expert is trained with its fixed configuration using the same frame-level scoring and video-level pooling formulation adopted at inference.

Algorithm 1 Training a visual expert E_m

Input: Visual dataset $\mathcal{D}_v$; expert E_m; classifier g_m; frame sampler $\mathcal{T}$; pooling rule $\mathcal{P}$
Output: Trained expert parameters Θ_m

for each training epoch **do**
  for mini-batch $\mathcal{B} \subset \mathcal{D}_v$ **do**
    for all $(V_i, y_v^{(i)}) \in \mathcal{B}$ **do**
      $\{\tilde{v}_{i,t}\}_{t=1}^{T} \leftarrow \mathcal{T}(V_i)$
      $q_{i,t,m} \leftarrow \sigma(g_m(E_m(\tilde{v}_{i,t})))$ for $t = 1, \ldots, T$
      $q_{i,m} \leftarrow \mathcal{P}(\{q_{i,t,m}\}_{t=1}^{T})$
    end for
    $\mathcal{L}_v^{(m)} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathrm{BCE}(q_{i,m}, \tilde{y}_v^{(i)})$
    Update Θ_m using $\nabla_{\Theta_m} \mathcal{L}_v^{(m)}$
  end for
end for
return Θ_m

The compact ensemble is intended to preserve complementary visual predictions while maintaining a fixed and analyzable inference structure. It should therefore be interpreted as a fixed ensemble, rather than a dynamically routed mixture-of-experts model.

3.4 Audio Expert Model

Audio manipulations may leave weak and temporally distributed traces in spectral structure, prosody, and longer-range acoustic continuity. The audio branch combines contextual speech representations with temporal sequence modeling.

Given a prepared waveform $\tilde{a}$, XLS-R extracts contextual acoustic features:

$$\mathbf{H} = \mathcal{F}_{\mathrm{XLS-R}}(\tilde{a}; \theta, \Delta\theta) \in \mathbb{R}^{L \times d_s}, \tag{13}$$

where L is the number of acoustic frames and d_s is the feature dimension. XLS-R is a self-supervised cross-lingual speech model built on wav2vec 2.0 pretraining [Babu *et al.*, 2021]. It provides contextual acoustic representations that can be adapted for audio authenticity prediction.

We adapt selected linear transformations with low-rank adaptation (LoRA) [Hu *et al.*, 2021]. For a frozen weight matrix $\mathbf{W} \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$, the adapted transformation is

$$\widetilde{\mathbf{W}} = \mathbf{W} + \frac{\lambda}{r}\mathbf{B}\mathbf{A}, \tag{14}$$

where $\mathbf{A} \in \mathbb{R}^{r \times d_{\mathrm{in}}}$ and $\mathbf{B} \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ are trainable low-rank matrices, r is the adaptation rank, and λ is a scaling factor.

The adapted features are projected into a task-specific sequence:

$$\mathbf{X} = \mathbf{H}\mathbf{W}_p + \mathbf{b}_p. \tag{15}$$

Mamba then contextualizes the sequence through selective state-space updates [Gu and Dao, 2023]. A simplified update is

$$\mathbf{h}_t = \overline{\mathbf{A}}_t \mathbf{h}_{t-1} + \overline{\mathbf{B}}_t \mathbf{x}_t, \qquad \mathbf{z}_t = \mathbf{C}_t \mathbf{h}_t + \mathbf{D}\mathbf{x}_t, \tag{16}$$

where $\overline{\mathbf{A}}_t$, $\overline{\mathbf{B}}_t$, and $\mathbf{C}_t$ depend on the current input. In SPARSEAVF, Mamba serves as a temporal aggregation component over contextualized XLS-R representations; it is not introduced as a new state-space formulation.

Algorithm 2 Training the audio expert

Input: Audio dataset $\mathcal{D}_a$; pretrained XLS-R $\mathcal{F}_{\mathrm{XLS-R}}$; LoRA parameters $\Delta\theta$
Output: Trained audio parameters Θ_a

Freeze the base XLS-R parameters θ
Initialize $\Theta_a = \{\Delta\theta, \mathbf{W}_p, \mathbf{b}_p, \Theta_{\mathrm{Mamba}}, \mathbf{W}_a, \mathbf{b}_a\}$
for each training epoch **do**
  for mini-batch $\mathcal{B} \subset \mathcal{D}_a$ **do**
    for all $(a_i, y_a^{(i)}) \in \mathcal{B}$ **do**
      $\tilde{a}_i \leftarrow \mathrm{PrepareAudio}(a_i)$
      $\mathbf{H}_i \leftarrow \mathcal{F}_{\mathrm{XLS-R}}(\tilde{a}_i; \theta, \Delta\theta)$
      $\mathbf{X}_i \leftarrow \mathbf{H}_i\mathbf{W}_p + \mathbf{b}_p$
      $\mathbf{Z}_i \leftarrow \mathrm{Mamba}(\mathbf{X}_i; \Theta_{\mathrm{Mamba}})$
      $\mathbf{u}_i \leftarrow [\mathrm{Mean}(\mathbf{Z}_i); \mathrm{Max}(\mathbf{Z}_i)]$
      $p_{a,i} \leftarrow \mathrm{softmax}(\mathbf{W}_a\mathbf{u}_i + \mathbf{b}_a)$
    end for
    $\mathcal{L}_a \leftarrow \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathrm{CE}(p_{a,i}, y_a^{(i)})$
    Update Θ_a using $\nabla_{\Theta_a}\mathcal{L}_a$
  end for
end for
return Θ_a

The resulting sequence $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_L]$ is summarized through mean and max pooling:

$$\mathbf{u} = \left[\frac{1}{L}\sum_{t=1}^{L}\mathbf{z}_t \; ; \; \max_{t=1,\ldots,L}\mathbf{z}_t\right]. \tag{17}$$

The audio authenticity posterior is

$$p_a(y_a \mid a) = \mathrm{softmax}(\mathbf{W}_a\mathbf{u} + \mathbf{b}_a), \qquad y_a \in \{R, F\}. \tag{18}$$

Audio expert training. Let $\mathcal{D}_a = \{(a_i, y_a^{(i)})\}_{i=1}^{N_a}$ denote the audio training set. The pretrained XLS-R backbone provides the initial acoustic representation, while the LoRA parameters, projection layer, Mamba module, and audio classifier are optimized for audio authenticity prediction.

3.5 Sparse Audio–Visual Fusion

The visual component of SPARSEAVF uses a compact preselected expert set, and every retained expert is evaluated for every input. The system does not use input-conditioned Top-K routing or dynamic expert selection. Its purpose is to maintain a bounded set of complementary visual predictors whose outputs remain explicit and analyzable.

After modality-specific prediction, the visual and audio branches produce posterior estimates for each paired sample. The fusion stage maps these modality-wise posteriors to the four joint authenticity states using a global log-linear formulation. It is intentionally low-capacity: it does not use raw feature concatenation, cross-attention, temporal synchronization, sample-specific modality quality, or input-dependent modality weighting.

Let $\alpha, \beta > 0$ denote global audio and visual weights, respectively, and let b_{y_v,y_a} denote a state-specific offset. For each joint state (y_v, y_a), the fusion score is

$$\begin{aligned} s_{y_v,y_a} = \beta \log\big(p_v(y_v \mid v) + \epsilon\big) \\ + \alpha \log\big(p_a(y_a \mid a) + \epsilon\big) + b_{y_v,y_a}, \end{aligned} \tag{19}$$

where ϵ is a small numerical constant. The joint posterior is

$$p(y_v, y_a \mid v, a) = \mathrm{softmax}\left(\{s_{y_v,y_a}\}_{y_v,y_a}\right), \tag{20}$$

and the binary manipulated probability is

$$p(y_{\mathrm{bin}} = 1 \mid v, a) = 1 - p(\mathrm{RR} \mid v, a). \tag{21}$$

When $\alpha = \beta = 1$ and all state-specific offsets are equal, the formulation reduces to a normalized product of the visual and audio posterior probabilities. The global weights and state-specific offsets therefore provide posterior reweighting rather than sample-adaptive cross-modal interaction.

Fusion optimization. The fusion head is trained with four-class cross-entropy:

$$\mathcal{L}_{\mathrm{fus}} = -\log p(y_v, y_a \mid v, a). \tag{22}$$

This formulation uses modality-wise posteriors as interpretable fusion inputs rather than an unconstrained concatenation of hidden representations. The fusion weights are shared across examples; therefore, the present model does not estimate sample-specific modality reliability or audiovisual correspondence.

4 Experiments

4.1 Benchmark, Task, and Official Metrics

We evaluate SPARSEAVF on the Track 2 benchmark for general AI-generated audio-video detection. The competition uses the released MVAD resources, whose repository identifies MVAD as the official dataset for the IJCAI 2026 DDL 2.0 Competition Track 2 [HuMengXue0104, 2026]. The benchmark covers realistic and anime-style content, as well as humans, animals, objects, and scenes. Each sample belongs to one of four modality-composition states: real visual–real audio (RR), fake visual–real audio (FR), real visual–fake audio (RF), or fake visual–fake audio (FF).

The current MVAD repository reports 205,758 multimodal video-audio samples generated by more than twenty generation methods [HuMengXue0104, 2026]. Its repository-wide composition contains 101,000 RR samples, 10,700 FR samples, 31,880 RF samples, and 62,178 FF samples. These repository-level counts indicate substantial class imbalance, but they should not be interpreted as the class distribution of the released competition package used for final evaluation.

The FR and RF states isolate visual and audio manipulation, respectively, whereas FF requires evidence from both streams. Binary detection therefore distinguishes RR from the union of FR, RF, and FF, while four-state recognition additionally requires modality attribution. The official Track 2 protocol reports AUC, accuracy, F1, and AP for binary and four-class tasks. Its final score combines binary and four-class scores with weights of 0.4 and 0.6, respectively [DDL 2.0 Organizing Committee, 2026].

Metric scope. We reproduce the aggregate values reported by the official scorer and retain its metric definitions and averaging conventions. The available official output does not provide per-class precision, recall, F1, AP, or a four-state confusion matrix. Consequently, the results below are not used to claim a specific improvement for RR, FR, RF, or FF individually.

Table 1: Benchmark and evaluation characteristics relevant to this study. Repository-scale statistics are cited from the released MVAD resource; audit coverage refers to the competition package analyzed in this work.

Property	Description
Benchmark resource	MVAD repository / DDL-GAV
Competition role	Official dataset for Track 2
Repository scale	205,758 audio-video samples
Generation coverage	More than 20 generation methods
Authenticity states	RR, FR, RF, and FF
Repository composition	101,000 RR; 10,700 FR; 31,880 RF; 62,178 FF
Audit coverage	11,557 released test videos
Evaluation tasks	Binary detection; four-state recognition
Official metrics	AUC, ACC, F1, and AP
Score weighting	0.4 binary + 0.6 four-class

Table 2: Final-pipeline configuration reported in this paper.

Component	Configuration
Visual input	384×384, 24 frames/video
Audio input	Mono, 16 kHz, 4-s segment
Visual branch	Compact fixed visual expert ensemble
Audio encoder	XLS-R with LoRA adaptation
Temporal audio model	Mamba
Visual aggregation	Fixed aggregation over retained experts
Fusion input	Visual and audio authenticity posteriors
Fusion form	Global log-linear posterior fusion
Fusion target	Four-state authenticity label
Model selection	Validation-set criterion

4.2 Implementation Configuration and Evaluation Scope

The final submission is one integrated audio-video detector. It combines a compact fixed visual expert ensemble, an XLS-R–LoRA–Mamba audio branch, and structured posterior fusion. The visual branch receives 24 sampled RGB frames at 384×384 resolution. The audio branch operates on mono 16-kHz waveforms and uses 4-s waveform segments.

The modality-specific branches are optimized using their corresponding binary authenticity labels. The fusion head is subsequently optimized using four-state labels, as described in Section 3. The final checkpoint is selected using a validation-set criterion. The official result is treated as a final competition evaluation outcome.

Scope of the reported comparison. The present paper reports the official result of the final integrated system. Development-stage configuration tests were used during system construction, but they do not constitute matched controlled ablations: their configurations do not isolate only one factor while holding the visual branch, audio branch, data split, optimization schedule, and checkpoint-selection procedure fixed. We therefore do not use such internal tests to attribute a numerical gain specifically to posterior fusion, LoRA adaptation, Mamba temporal modeling, or the visual expert ensemble.

Table 3: Official evaluation of the final single submitted model.

Task	**AUC**	**ACC**	**F1**	**AP**	**Score**
Binary	0.9480	0.9187	0.9184	0.9691	0.9396
Four-state	0.8810	0.8273	0.6726	0.7191	0.7746
				Final weighted score	**0.8406**

4.3 Descriptive Media-Quality Audit

The released competition package contains substantial media-quality variation in addition to authenticity variation. We conduct a descriptive no-reference audit over 11,557 released test videos. Figure 2 uses an active-label threshold of 20 and groups diagnostic severity into light (1–34), moderate (35–69), and severe (70–100). The indicators are non-exclusive: one video may activate multiple visual and audio diagnostics.

Compression is the most prevalent visual diagnostic (79.74%), followed by visual noise (44.22%) and resolution loss (29.51%). For audio, noise floor (76.94%) and low-pass loss (63.30%) are common, while mono conversion (45.44%) and downsampling (45.37%) occur almost entirely in the severe band. These measurements characterize media-quality variation in the released package; they do not establish that a sample is manipulated, nor do they validate the diagnostic categories as ground-truth degradation labels.

The audit is used only for post-hoc description. It does not provide training labels, validation criteria, fusion parameters, threshold selection, model-selection signals, or test-time adaptation. In particular, `splice_suspected` denotes a possible splice or abnormal transition identified by the diagnostic procedure and is not a manual splice annotation. Since no model accuracy is reported within the audit strata, the audit should not be interpreted as a direct robustness evaluation.

4.4 Official Evaluation of the Final Submitted Model

The competition permits one submitted model per track. Accordingly, the following results correspond to the single integrated detector described in Section 3. The official scorer reports 0.9480 AUC, 0.9187 accuracy, 0.9184 F1, and 0.9691 AP for binary detection. For four-state recognition, it reports 0.8810 AUC, 0.8273 accuracy, 0.6726 F1, and 0.7191 AP.

Interpretation of the official result. The final submitted system obtains a weighted official score of 0.8406. The four-state score contributes more strongly to this objective because of the organizer-defined $0.4/0.6$ binary/four-state weighting. The reported four-state result is consistent with the intended role of structured posterior fusion: the system first estimates visual and audio authenticity separately, then maps their combination to the joint state space.

This interpretation is limited in scope. The official output is aggregate and does not identify which state contributes most to the four-state result. It also does not establish that the global fusion weights model sample-dependent audio-video

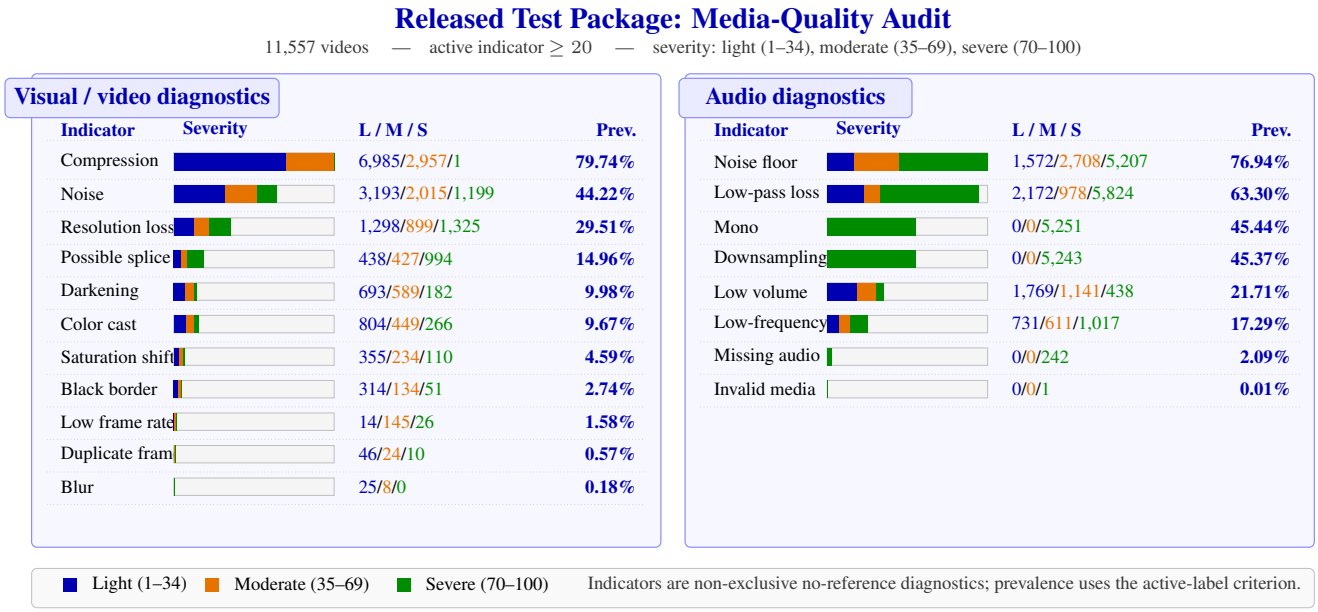

Figure 2: Descriptive no-reference media-quality audit over the released competition package (11,557 videos). Bars show light, moderate, and severe diagnostic counts; the rightmost column gives prevalence under an active-label threshold of 20. Indicators are non-exclusive and are not manual degradation or authenticity labels.

correspondence, modality quality, or reliability. Similarly, the descriptive media-quality audit does not establish robustness under compression, noise, resampling, or other degradation conditions. These questions require matched ablation, per-class, and degradation-stratified evaluations beyond the aggregate official output reported here.

5 Conclusion

This work formulates AI-generated audio-video detection as structured joint prediction over visual and audio authenticity. The formulation explicitly represents the real–real, fake–real, real–fake, and fake–fake states, thereby preserving modality-wise semantics when manipulation affects only one stream.

The proposed single integrated detector combines a compact fixed visual expert ensemble with an XLS-R–LoRA–Mamba audio branch. It first produces modality-wise authenticity posteriors and then maps them to the four-state label space through a global log-linear posterior fusion head. This design provides an interpretable separation between visual evidence, audio evidence, and joint-state prediction, while avoiding the assumption that all manipulated samples exhibit the same audio-video composition.

Under the official Track 2 evaluation, the final submitted model achieves a binary score of 0.9396, a four-state score of 0.7746, and a weighted score of 0.8406. We additionally provide a descriptive media-quality audit of the released evaluation package to characterize variation in visual and acoustic quality indicators. This audit is not used for training, model selection, or direct robustness evaluation.

The current framework has several limitations. The visual expert set is fixed and does not perform input-conditioned routing. The fusion head uses globally shared modality weights and state-specific offsets; it does not model sample-specific modality reliability, audiovisual correspondence, or synchronization. Moreover, the available official output provides aggregate four-state metrics rather than per-class results, so it does not support state-specific conclusions for FR or RF. Future work should examine controlled fusion ablations, out-of-fold fusion training, per-state evaluation, robustness under controlled media degradation, and generalization to unseen generators, languages, and real-world transmission conditions.

References

[Babu *et al.*, 2021] 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. *arXiv preprint arXiv:2111.09296*, 2021.

[Cai *et al.*, 2023] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. *arXiv preprint arXiv:2311.15308*, 2023.

[DDL 2.0 Organizing Committee, 2026] DDL 2.0 Organizing Committee. Track 2: General AIGC audio-video detection. https://ai-safety-workshop-ijcai2026.github.io/Track2.html, 2026. Accessed: 2026-07-01.

[Durall *et al.*, 2020] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spec-

tral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[Frank *et al.*, 2020] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Hu *et al.*, 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[HuMengXue0104, 2026] HuMengXue0104. MVAD: A benchmark dataset for multimodal ai-generated video-audio detection. https://github.com/HuMengXue0104/MVAD, 2026. GitHub repository. Accessed: 2026-07-02.

[Jung *et al.*, 2021] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. *arXiv preprint arXiv:2110.01200*, 2021.

[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. *arXiv preprint arXiv:2109.02993*, 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.

[Ojha *et al.*, 2023] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

[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. *arXiv preprint arXiv:2406.02951*, 2024.

[Wang *et al.*, 2020] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.

[Wang *et al.*, 2023] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[Yamagishi *et al.*, 2021] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.