# Modality-Specific Latent Realness Modeling for General AI-Generated Audio-Video Detection

**Jianfeng Dong**[1,2] , **Haoran Chen**[1] , **Daiqi Gao**[1] , **Tianzheng Wang**[1] ,
**Mingming Deng**[1] , **Yirong Sun**[3] and **Xun Wang**[1,2*]
[1]Zhejiang Gongshang University, Hangzhou, China
[2]Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology, Hangzhou, China
[3]University of Nottingham Ningbo China, Ningbo, China
[*]Corresponding author

## Abstract

General AI-generated audio-video detection requires determining whether the visual stream, the audio stream, or both modalities are synthetic. In this paper, we propose DiFuse, a modality-specific latent realness modeling framework for the IJCAI 2026 DDL-GAV challenge. DiFuse uses frozen pretrained audio and visual encoders to extract modality-level representations, learns authenticity cues in each latent space, and fuses the resulting evidence for four-class audio-video prediction. For audio detection, we build a Dasheng-based branch with duration-aware cropping and temporal mean–std pooling. For video detection, we extract DINOv3 frame features and introduce a class-conditional feature shuffling strategy to encourage sparse forgery evidence learning before attention-based aggregation. On the official test set, DiFuse achieves a final score of 0.8438, with a binary score of 0.9395 and a four-class score of 0.78, demonstrating the effectiveness of modality-specific realness modeling and audio-video fusion for general AIGC audio-video detection.

## 1 Introduction

Recent advances in generative AI have substantially improved the realism, accessibility, and diversity of synthetic multimedia content. Text-to-video systems can synthesize temporally coherent visual streams [Brooks *et al.*, 2024; Bar-Tal *et al.*, 2024; Yang *et al.*, 2025], while modern speech, audio, and audio-video generation models can produce plausible acoustic content for a wide range of scenes [Wang *et al.*, 2023a; Liu *et al.*, 2023; Polyak *et al.*, 2024]. As a result, multimedia forensics is no longer limited to visual authenticity detection. In real-world audio-video samples, the visual stream and the audio stream may be independently authentic or synthetic, or both may be AI-generated. This emerging threat landscape requires detection systems to assess modality-specific authenticity while also reasoning about cross-modal consistency.

This need for joint reasoning exposes the limitations of conventional face-centric or speech-centric deepfake detection. Earlier forensic tasks often rely on face swapping artifacts, speaker spoofing cues, or lip-sync errors [Jung *et al.*, 2022; Li *et al.*, 2020; Haliassos *et al.*, 2021], and recent visual detectors further improve robustness by modeling localized or temporal artifacts [Nguyen *et al.*, 2024; Nguyen *et al.*, 2025]. These assumptions are valuable for talking-face manipulation, but they do not fully cover contemporary AIGC media, which may contain open-domain scenes, anime-style videos, music, ambient sounds, and complex audio-visual events. Moreover, synthetic content does not necessarily cause an obvious cross-modal mismatch: generated audio can be compatible with real video, and jointly generated audio-video samples can remain plausible in both semantics and timing. Therefore, audio-visual AIGC detection should move beyond face-centric and mismatch-based cues toward joint modeling of modality-specific authenticity and cross-modal consistency. Consistent with this motivation, the frequency spectra in Figure 1 show that real and AI-generated samples exhibit distinguishable frequency-domain patterns in both visual and audio streams, suggesting that modality-specific artifacts remain informative even when cross-modal mismatches are not explicit.

The DDL-GAV General AIGC Audio-Video Detection Challenge is designed around this broader audio-visual setting. Built on MVAD [Hu *et al.*, 2026], it covers diverse content categories, visual styles, and generation techniques, rather than focusing only on talking faces, celebrity identities, or speech-driven forgery. Specifically, DDL-GAV defines four modality combinations: **real-real**, **fake-real**, **real-fake**, and **fake-fake**. This formulation requires modality-specific evidence, as neither audio-only nor video-only detectors can handle all forgery types.

Prior work provides useful foundations, but its assumptions only partially fit DDL-GAV. Audio spoofing detectors mainly study synthetic or converted speech [Jung *et al.*, 2022; Tak *et al.*, 2022], visual deepfake detectors often focus on face-region, frequency, or temporal artifacts [Li *et al.*, 2020; Liu *et al.*, 2021; Zheng *et al.*, 2021], and many audio-video datasets or methods emphasize identity, synchronization, temporal localization, or cross-modal inconsistency [Khalid *et al.*, 2021; Cai *et al.*, 2023; Cai *et al.*, 2024; Katamneni and Rattani, 2024]. Yet DDL-GAV covers open-domain visuals and diverse audio types, where these task-specific cues may

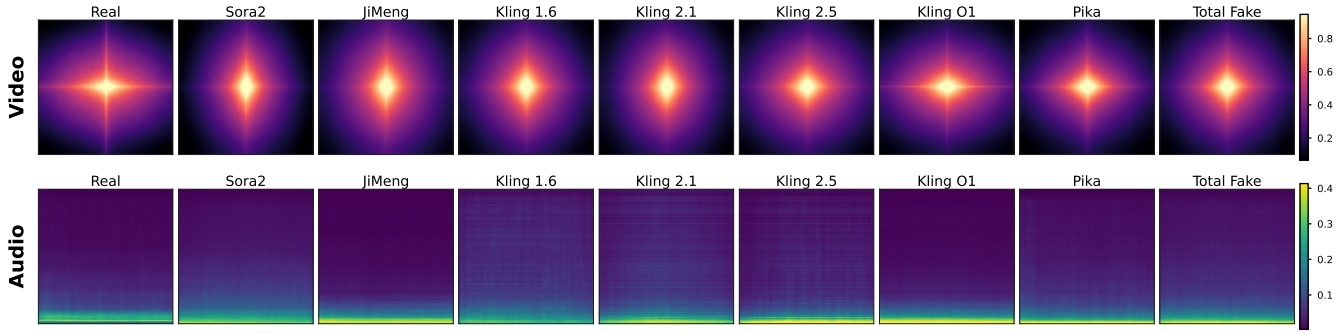


Figure 1: Video and audio frequency spectra for MVAD Dataset. Brighter regions indicate stronger frequency energy.

be unreliable. Moreover, an overall real/fake prediction may hide whether the visual stream, the audio stream, or both are AI-generated.

We model task-specific real support in pretrained latent spaces. Despite variations in content, style, and modality conditions, strong foundation encoders can map real audio and video onto relatively stable manifolds, enabling real samples to serve as references without explicitly modeling the full real-world distribution. Specifically, we use frozen DINOv3-7B [Siméoni *et al.*, 2025] and Dasheng-1.2B [Dinkel *et al.*, 2024] as visual and audio encoders. Rather than acting as forensic detectors, they provide modality-specific latent spaces in which authenticity boundaries are learned. We further introduce *class-conditional feature shuffling*, which applies controlled perturbations to real features to approximate their local neighborhoods. Attention pooling aggregates evidence from the original and augmented features, producing modality-level authenticity scores, which are finally combined through cross-modal fusion.

In summary, we address DDL-GAV as a general audio-video AIGC detection problem that requires modality-specific authenticity modeling beyond face-centric, speech-centric, and mismatch-based cues. We propose class-conditional feature shuffling, which approximates local real neighborhoods in pretrained audio and visual latent spaces to learn modality-level authenticity scores.

## 2 Related Work

### 2.1 Synthetic Audio and Video Generation

Recent generative models have substantially expanded the realism and diversity of synthetic visual content. Large-scale text-to-video and image-to-video systems, including latent video diffusion, Lumiere, Sora, and CogVideoX, can generate temporally coherent videos covering diverse scenes and visual styles [Blattmann *et al.*, 2023; Bar-Tal *et al.*, 2024; Brooks *et al.*, 2024; Yang *et al.*, 2025]. Consequently, the visual stream of an online video can no longer be assumed to be authentic.

Audio generation has progressed in parallel. VALL-E shows that neural codec language models can synthesize speech from a short acoustic prompt [Wang *et al.*, 2023a], while AudioLDM and AudioLDM 2 extend latent diffusion and self-supervised pretraining to general text-to-audio generation [Liu *et al.*, 2023; Liu *et al.*, 2024]. MusicGen further demonstrates controllable generation of non-speech music content [Copet *et al.*, 2023]. As a result, forged audio in AIGC videos is not limited to cloned voices; it may also appear as generated music, action sounds, background ambience, or other non-speech acoustic events.

More importantly, recent systems increasingly generate audio and video jointly or condition one modality on the other. FoleyCrafter synthesizes video-synchronized sounds from silent videos [Zhang *et al.*, 2026], and Movie Gen integrates video generation, editing, personalization, and synchronized audio generation in a unified media generation framework [Polyak *et al.*, 2024]. This trend weakens the traditional assumption that fake audio-video content must contain obvious cross-modal mismatch: generated audio may be semantically and temporally compatible with the visual stream, and both modalities may be synthesized together.

### 2.2 AI-Generated Audio-Video Detection

AI-generated audio-video detection builds on three related lines of work: audio forgery detection, visual forgery detection, and multimodal audio-video detection. Existing audio forgery detectors cover both speech-oriented anti-spoofing and general-audio detection. AASIST models spectro-temporal relations using heterogeneous graph attention [Jung *et al.*, 2022], while RawBMamba directly processes raw waveforms and employs bidirectional state-space modeling to capture both short- and long-range forgery artifacts [Chen *et al.*, 2024]. Large audio encoders provide another direction. XLS-R learns cross-lingual speech representations [Babu *et al.*, 2022], whereas BEATs learns general acoustic representations through acoustic tokenizers [Chen *et al.*, 2023]. For non-speech audio, Ouajdi et al. combine CLAP embeddings with an MLP to detect synthetic environmental and Foley sounds [Ouajdi *et al.*, 2024]. These general-audio representations are particularly relevant to MVAD, whose audio tracks contain not only speech but also music, sound effects, actions, and environmental ambience. Following this direction, our audio branch combines frozen Dasheng-1.2B representations with a lightweight MLP classifier [Dinkel *et al.*, 2024].

Visual forgery detection has historically focused on face manipulation and synthetic video artifacts. Face X-Ray detects blending boundaries between source images [Li *et al.*, 2020], SPSL exploits frequency traces [Liu *et al.*, 2021], and

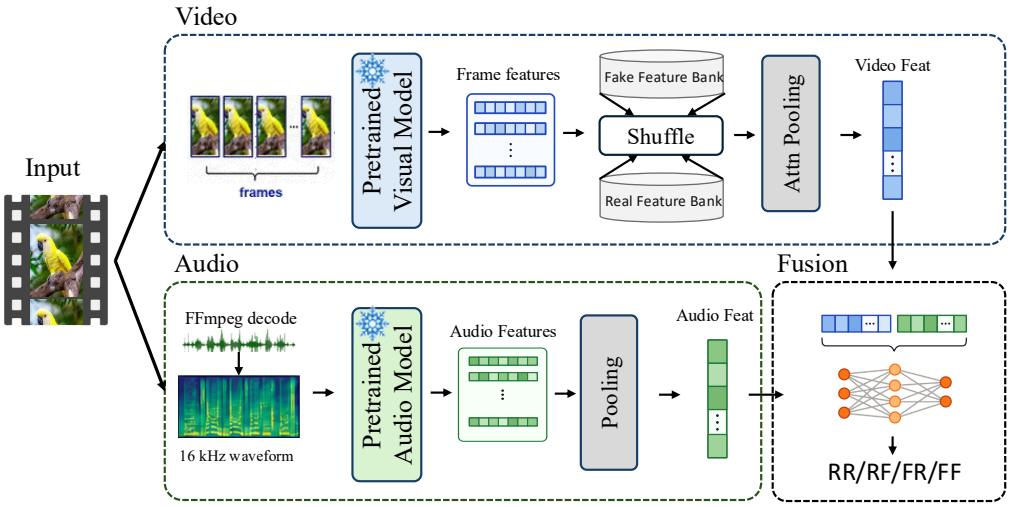


Figure 2: Overview of the proposed DiFuse framework for general AI-generated audio-video detection.

FTCN or AltFreezing model temporal or spatio-temporal artifacts [Zheng *et al.*, 2021; Wang *et al.*, 2023b]. Recent detectors such as LAA-Net and FakeSTormer explicitly model localized, subtle, or spatio-temporal artifacts for stronger generalization [Nguyen *et al.*, 2024; Nguyen *et al.*, 2025]. These methods provide strong visual evidence, but they cannot detect samples where only the audio stream is fake.

Audio-video deepfake detection further models the relationship between modalities. Datasets such as FakeAVCeleb, LAV-DF, and AV-Deepfake1M provide multimodal labels, temporal localization, or large-scale audio-visual forgery settings [Khalid *et al.*, 2021; Cai *et al.*, 2023; Cai *et al.*, 2024]. Joint audio-visual detection and AVoiD-DF exploit cross-modal attention or inconsistency cues [Zhou and Lim, 2021; Yang *et al.*, 2023]. Recent multimodal methods also revisit contrastive learning and contextual cross-modal modeling for audio-visual detection and localization [Zhang *et al.*, 2023; Katamneni and Rattani, 2024]. However, many of these works remain close to talking-face or mismatch-based settings. MVAD and the DDL-GAV challenge extend the task to general AIGC audio-video detection with four modality combinations: real-real, fake-real, real-fake, and fake–fake [Hu *et al.*, 2026]. This setting requires preserving modality-specific authenticity evidence while using cross-modal cues as complementary information.

# 3 Method

## 3.1 Overall Architecture

We propose *DiFuse*, a divide-then-fuse framework for General AIGC Audio-Video Detection based on modality-specific latent realness modeling. This formulation estimates the authenticity of each modality within frozen pretrained latent spaces, motivated by the observation that real audio and visual signals form relatively stable manifolds in well-pretrained representation spaces, whereas generated samples tend to deviate through spectral, temporal, or visual artifacts. As shown in Figure 2, DiFuse maps video frames and audio tracks into separate visual and audio representation spaces using large-scale pretrained models as frozen backbones. It then learns realness cues with modality-specific audio and visual branches by modeling deviations from real samples, and finally fuses the resulting evidence to predict the four-class audio-video authenticity label.

## 3.2 Audio Branch

As illustrated in the lower part of Figure 2, the audio branch extracts modality-specific evidence from the input video. For each video, we first use FFmpeg to decode the audio track into a 16-kHz mono waveform. The waveform is then divided into duration-aware 6-second crops: audio shorter than 6 seconds is padded to form one crop, audio between 6 and 12 seconds is divided into two crops, and audio longer than 12 seconds is divided into three crops.

We use Dasheng-1.2B [Dinkel *et al.*, 2024] as the frozen pretrained audio encoder. Its general audio pretraining is well suited to MVAD, since the audio tracks contain diverse acoustic events, including speech, music, sound effects, actions, and environmental sounds. For the $k$-th audio crop, Dasheng produces temporal hidden states $\mathbf{Z}_k \in \mathbb{R}^{T_k \times 1536}$, where $T_k$ denotes the number of encoder time tokens.

Let $\mathbf{Z}_k = [\mathbf{z}_{k,1}, \ldots, \mathbf{z}_{k,T_k}]$ denote the sequence of Dasheng temporal hidden states from crop $k$. Following the pooling module shown in Figure 2, we summarize each crop by concatenating the temporal mean and standard deviation:

$$\boldsymbol{\mu}_k = \frac{1}{T_k}\sum_{t=1}^{T_k} \mathbf{z}_{k,t}, \quad \boldsymbol{\sigma}_k = \sqrt{\frac{1}{T_k}\sum_{t=1}^{T_k} (\mathbf{z}_{k,t} - \boldsymbol{\mu}_k)^2},$$

$$\mathbf{h}_k = [\boldsymbol{\mu}_k; \boldsymbol{\sigma}_k] \in \mathbb{R}^{3072}.$$

Here $[\cdot;\cdot]$ denotes vector concatenation. For Dasheng, this operation produces a 3072-dimensional crop-level feature. The crop features are cached on disk and used as the basic audio representation for downstream detection.

For audio-video fusion, we further aggregate the crop-level

features into a video-level audio representation:

$$z_a = \frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_k,$$

where $K$ is the number of audio crops. The resulting $z_a \in \mathbb{R}^{3072}$ is used as the audio input to the fusion module.

For audio-only detection, the cached crop features are fed into a compact MLP classifier with LayerNorm, two hidden layers $3072 \to 1024 \to 256$, GELU activations, dropout, and a single output logit. For a video with $K$ audio crops, the video-level audio fake probability is obtained by averaging crop-level logits:

$$p_a(\text{fake} \mid \mathbf{v}) = \sigma\left(\frac{1}{K}\sum_{k=1}^{K} g(\mathbf{h}_k)\right),$$

where $g(\cdot)$ denotes the MLP logit function and $\sigma(\cdot)$ is the sigmoid function.

To examine whether general-purpose audio representations are more suitable for MVAD than speech-specific or anti-spoofing representations, we compare Dasheng-1.2B with speech-oriented XLS-R features [Babu *et al.*, 2022], the end-to-end AASIST anti-spoofing model [Jung *et al.*, 2022], and a BEATs/UniLM baseline [Chen *et al.*, 2023].

### 3.3 Video Branch

As illustrated in the upper part of Figure 2, the video branch uniformly samples $N_v$ frames and extracts frame-level features using a frozen pretrained visual encoder. Since forgery cues in fake videos are often sparse and unevenly distributed, only a small subset of frames may contain clear artifacts, while many others appear realistic. Therefore, the detector should identify discriminative frames from the sampled sequence rather than rely on uniform temporal aggregation.

**Realness-Guided Latent Feature Shuffling.** Inspired by [Chen *et al.*, 2026], we implement latent realness modeling in the visual branch by directly shuffling features within the pretrained latent space. Within each mini-batch, frame-level features are divided into an in-batch real feature bank $\mathcal{B}_r$ and an in-batch fake feature bank $\mathcal{B}_f$ according to their video labels. The two banks are reconstructed from the current mini-batch at every training iteration and are not maintained across different batches.

Real features are kept unchanged and serve as latent realness references. For each fake feature $h_f \in \mathcal{B}_f$, we randomly sample a real feature $h_r \in \mathcal{B}_r$ from the same mini-batch and construct

$$\widetilde{h}_f = (1-r)h_f + rh_r,$$

where $r \in [0,1]$ is the shuffle ratio controlling the amount of realness evidence injected into the fake feature. A larger $r$ introduces a greater contribution from the randomly paired real feature, while $r = 0$ leaves the original fake feature unchanged. Here, feature shuffling refers to the random in-batch pairing and weighted mixing of fake and real latent features. This one-sided operation injects realistic evidence into fake representations while retaining their fake labels, thereby constructing harder samples with weaker forgery cues and encouraging the detector to learn a more discriminative latent realness boundary.

Table 1: First-frame linear probing comparison of different visual backbones. We extract features from the first frame of each video using a frozen visual encoder and train a linear classifier to evaluate backbone representation quality for video forgery detection.

| **Method** | **Acc** | **AUROC** | **AP** | **Fake-F1** |
|---|---|---|---|---|
| D3 [Zheng *et al.*, 2025] | 0.7220 | 0.5151 | 0.1570 | 0.2339 |
| DINOv2 Large | 0.7534 | 0.9025 | 0.5727 | 0.4954 |
| DINOv3 ViT-B/16 | 0.7862 | 0.9414 | 0.7653 | 0.5383 |
| DINOv3 ConvNeXt-L | 0.8522 | 0.9347 | 0.6747 | 0.6139 |
| DINOv3 ViT-7B/16 | 0.8637 | 0.9561 | 0.8423 | 0.6421 |

Table 2: Validation performance of audio forgery detection methods on the 9:1 MVAD split. Results are selected by the best validation ROC-AUC.

| Method | ROC-AUC | AP | Acc. | EER |
|---|---|---|---|---|
| *Existing audio forgery methods* | | | | |
| AASIST [Jung *et al.*, 2022] | 0.99788 | 0.99346 | 0.97981 | 0.01937 |
| BEATs [Chen *et al.*, 2023] | 0.99852 | 0.99636 | 0.98687 | 0.01498 |
| CLAP-MLP [Ouajdi *et al.*, 2024] | 0.99781 | 0.99447 | 0.97995 | 0.01919 |
| RawBMamba [Chen *et al.*, 2024] | 0.99301 | 0.97637 | 0.96392 | 0.03630 |
| *Encoder+MLP comparison* | | | | |
| XLS-R+MLP [Babu *et al.*, 2022] | <u>0.99937</u> | <u>0.99815</u> | <u>0.99033</u> | <u>0.00911</u> |
| Ours: Dasheng-1.2B+MLP [Dinkel *et al.*, 2024] | **0.99987** | **0.99961** | **0.99605** | **0.00379** |

**Attention-Based Pooling.** Let

$$\widetilde{\mathbf{H}}_v = [\widetilde{h}_1, \widetilde{h}_2, \ldots, \widetilde{h}_{N_v}] \in \mathbb{R}^{N_v \times d_v}$$

denote the frame-level feature sequence after latent feature shuffling, where features from real videos remain unchanged. We use a learnable attention pooling module to aggregate the frame-level evidence. For each feature $\widetilde{h}_i$, the attention score is computed as

$$s_i = w_2^\top \tanh(W_1\widetilde{h}_i + b_1) + b_2,$$

where $W_1 \in \mathbb{R}^{256\times d_v}$ and $w_2 \in \mathbb{R}^{256}$. The normalized attention weight is given by

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{N_v}\exp(s_j)}, \qquad \sum_{i=1}^{N_v}\alpha_i = 1.$$

The final video-level representation is obtained as

$$z_v = \sum_{i=1}^{N_v}\alpha_i\widetilde{h}_i.$$

Compared with average pooling, attention pooling acts as a soft evidence selector, assigning larger weights to frames with stronger forgery-related cues and reducing the dilution caused by realistic-looking or less informative frames.

### 3.4 Audio-Video Detection

The audio-video detector fuses modality-specific evidence from the audio and video branches. Let $z_v \in \mathbb{R}^{d_v}$ and $z_a \in \mathbb{R}^{d_a}$ denote the video and audio representations, respectively. We concatenate them and feed the result into a lightweight multimodal projector:

$$z_{av} = [z_v; z_a], \quad \tilde{z}_{av} = f_\theta(z_{av}),$$

Table 3: Audio-visual fusion results on the validation set using DaSheng and DINOv3-ViT-7B as audio and video backbones.

| Fusion Setting | Acc. | AUC | AP | Fake F1 | Real F1 |
|---|---|---|---|---|---|
| Linear Probing | 0.9888 | 0.9993 | 0.9967 | 0.9598 | 0.9935 |
| w/o Shuffle | 0.9950 | 0.9999 | 0.9993 | 0.9811 | 0.9971 |
| w/ Shuffle | 0.9969 | 0.9999 | 0.9998 | 0.9941 | 0.9979 |

Table 4: MVAD partition used for validation in the four-class audio-video detection task.

| Split | Total | RR | FR | RF | FF |
|---|---|---|---|---|---|
| Train | 127,504 | 87,661 | 6,240 | 22,681 | 10,922 |
| Validation | 14,164 | 9,740 | 693 | 2,519 | 1,212 |

where $f_\theta(\cdot)$ denotes the projector.

The main classifier predicts four audio-video authenticity categories:

$$\mathcal{Y} = \{\textit{real-real}, \textit{real-fake}, \textit{fake-real}, \textit{fake-fake}\},$$

where the first term indicates the video modality and the second term indicates the audio modality. The four-class probability distribution is computed as

$$p = \mathrm{softmax}(W\tilde{z}_{av} + b).$$

To preserve modality-specific evidence during fusion, we attach three auxiliary binary heads for any-fake, video-fake, and audio-fake prediction. The any-fake head estimates whether at least one modality is manipulated, while the video-fake and audio-fake heads provide modality-specific supervision. The overall training objective is

$$\mathcal{L} = \mathcal{L}_4 + \lambda\left(\mathcal{L}_{any} + \mathcal{L}_{video} + \mathcal{L}_{audio}\right),$$

where $\mathcal{L}_4$ denotes the primary four-class classification loss, and $\mathcal{L}_{any}$, $\mathcal{L}_{video}$, and $\mathcal{L}_{audio}$ denote the three auxiliary binary classification losses. We set the common auxiliary-loss coefficient to $\lambda = 0.35$ in all experiments. This multi-task supervision encourages the fused representation to retain both joint audio-video evidence and modality-specific forgery cues.

# 4 Experiments

## 4.1 Experiment Setup

**Dataset.** We evaluate the modality-specific branches on the MVAD training split used by the DDL-GAV challenge. Following the four labels, RR, FR, RF, and FF correspond to $(y_v, y_a) = (0,0), (1,0), (0,1), (1,1)$, where 0 and 1 denote real and fake, respectively. We use a 9:1 train/validation partition over all available generator groups with random seed 42. Table 4 summarizes the resulting split.

**Evaluation Metrics.** Following the validation-style reporting used in prior IJCAI workshop challenge papers, we report validation metrics rather than hidden test labels. The primary metric is ROC-AUC. We additionally report average precision (AP), accuracy, and equal error rate (EER). ROC-AUC and AP measure ranking quality under the imbalanced validation distribution, while accuracy and EER summarize thresholded detection behavior.

Table 5: Ablation of the feature shuffle ratio on a manually annotated OOD subset randomly sampled from MVAD ($N = 673$). All variants use the same backbone and differ only in the shuffle ratio $r$. RR, RF, FR, and FF denote real-real, real-fake, fake-real, and fake-fake, respectively.

| | Overall | | Class Recall | | | |
|---|---|---|---|---|---|---|
| Method | Acc.↑ | Macro-F1↑ | RR↑ | RF↑ | FR↑ | FF↑ |
| No shuffle | 0.6597 | 0.4982 | 0.5842 | 0.9394 | 0.0000 | 0.8582 |
| $r = 0.2$ | 0.7786 | **0.5983** | 0.7943 | 0.9545 | 0.2500 | 0.7015 |
| $r = 0.4$ | **0.7920** | 0.5921 | 0.8249 | 0.9545 | 0.1875 | 0.6716 |
| $r = 0.6$ | 0.7801 | 0.5842 | 0.8249 | 0.9545 | 0.2500 | 0.6045 |
| $r = 0.8$ | 0.7474 | 0.5765 | 0.7593 | 0.9545 | 0.2500 | 0.6642 |

**Implementation Details.** We freeze Dasheng-1.2B and DINOv3 ViT-7B/16 and train prediction heads on cached features. Audio features are mean-pooled to 3072 dimensions and projected to 512 dimensions. For video, 4096-dimensional features from 16 uniformly sampled frames are attention-pooled and projected to 512 dimensions. The two representations are concatenated for four-class prediction, with auxiliary audio-fake, video-fake, and any-fake heads. We set $\lambda = 0.35$ and train with AdamW for 6 epochs using a learning rate of $10^{-4}$, weight decay of $10^{-4}$, batch size 2048, and dropout 0.25. At inference, $P(\text{fake}) = 1 - P(\texttt{real\_real})$.

## 4.2 Experimental Results and Analysis

Figure 1 shows clear spectral differences between real and AI-generated samples in both video and audio streams. Fake samples exhibit distinct energy distributions and frequency textures across low-, mid-, and high-frequency regions, suggesting that frequency-domain cues provide useful evidence for audio-visual forgery detection.

Table 2 compares the proposed audio branch with four representative audio forgery detectors and an XLS-R+MLP counterpart on the same 9:1 MVAD split. Among the baselines, BEATs performs best, achieving 0.99852 ROC-AUC, 0.99636 AP, 0.98687 accuracy, and 0.01498 EER. AASIST and RawBMamba, mainly designed for synthetic or converted speech, perform less effectively on MVAD. CLAP-MLP, despite pretraining on environmental and Foley audio, reaches only 0.99781 ROC-AUC and 0.97995 accuracy, suggesting that neither speech-specific cues nor general environmental-audio representations fully capture the diverse audio content in MVAD.

Dasheng-1.2B+MLP achieves the best overall results, with 0.99987 ROC-AUC, 0.99961 AP, 0.99605 accuracy, and 0.00379 EER. Under the controlled Encoder+MLP setting, it improves accuracy over XLS-R from 0.99033 to 0.99605 and reduces EER from 0.00911 to 0.00379. It also outperforms all other methods across the four metrics, supporting the use of a general-purpose audio encoder for MVAD, which contains synthetic speech, music, sound effects, and environmental audio generated by heterogeneous AIGC pipelines.

Table 1 shows that stronger visual representations are crucial for video forgery detection. D3 achieves only 0.7220 accuracy and 0.5151 AUROC, while pretrained encoders pro-

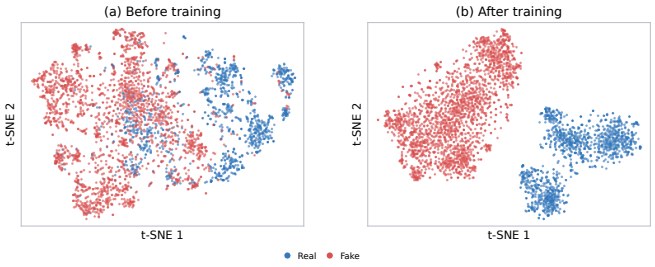


Figure 3: t-SNE visualization of visual features before and after training. Blue and red points denote real and fake samples, respectively.

vide substantial gains. DINOv2-Large reaches 0.9025 AUROC, and DINOv3 performs even better. DINOv3 ViT-7B/16 achieves the best results, with 0.8637 accuracy, 0.9561 AUROC, 0.8423 AP, and 0.6421 fake F1. However, the relatively low fake F1 indicates that video-only detection remains challenging, likely because visual artifacts are sparse, subtle, and unevenly distributed across frames.

Table 3 evaluates different audio-visual fusion settings using DaSheng and DINOv3-ViT-7B as the audio and video backbones. All variants achieve strong performance, with AUC and AP close to 1.0, confirming the complementarity of audio and visual evidence. Compared with linear probing, the learned fusion variants improve fake F1 from 0.9598 to 0.9811 and 0.9941, showing that trainable fusion better captures forgery-related cues. The best setting, w/ Shuffle, achieves 0.9969 accuracy, 0.9999 AUC, 0.9998 AP, 0.9941 fake F1, and 0.9979 real F1, demonstrating the effectiveness of joint audio-visual modeling for MVAD detection.

Different models and variants often achieve similarly high performance on the in-domain validation set, making their differences difficult to distinguish. We therefore train on the training set, select checkpoints on the validation set, and evaluate them on a small manually annotated OOD subset of 673 randomly sampled MVAD examples. As shown in Table 5, $r = 0.4$ improves accuracy from 0.6597 to 0.7920, demonstrating the effectiveness of feature shuffling.

We adapt AVFF [Oorloff *et al.*, 2024] to the four-class setting by replacing its binary head with a four-way classifier and fine-tuning the backbone for one epoch on the same in-distribution split. AVFF achieves strong performance on the MVAD validation set, reaching 0.9864 accuracy and 0.9755 Macro-F1. However, its performance drops sharply on the manually annotated OOD set, achieving only 0.1917 accuracy and 0.1540 Macro-F1, which reveals limited out-of-domain generalization.

### 4.3 Visualization

Figure 3 shows that pretrained visual features already provide real-fake separability, while detector training further improves class compactness and discrimination, indicating that frozen visual representations contain useful forgery cues. Figure 4 presents audio feature distributions from different pretrained encoders. Without forgery-specific training, these representations still exhibit real-fake separation, especially for Dasheng-1.2B, suggesting that large-scale pretraining captures natural audio patterns and that lightweight classifiers can effectively adapt them for audio-visual forgery detection.

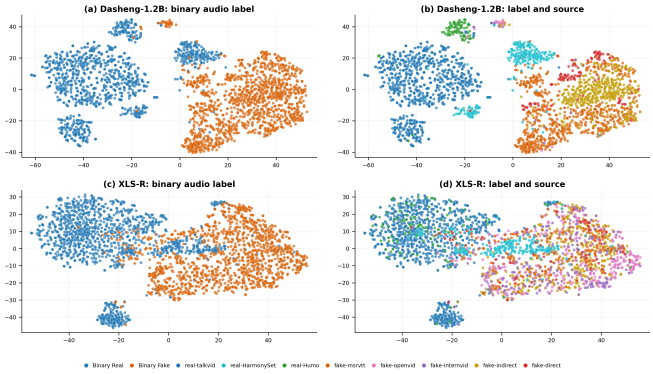


Figure 4: t-SNE visualization of audio embeddings from Dasheng-1.2B and XLS-R. Dasheng-1.2B achieves clearer real-fake separation while preserving generation-source clusters, indicating more discriminative audio representations.

## 5 Conclusion

In this paper, we presented DiFuse, a modality-specific latent realness modeling framework for general AI-generated audio-video detection. DiFuse separately models audio and video authenticity using frozen Dasheng and DINOv3 representations, with duration-aware audio pooling and class-conditional feature shuffling for sparse visual evidence learning. The modality-specific evidence is then fused for four-class prediction. On the official DDL-GAV test set, DiFuse achieves a final score of 0.8438, with a binary score of 0.9395 and a four-class score of 0.78, showing the effectiveness of modality-specific modeling and audio-video fusion. A limitation is that the current evaluation is conducted only on MVAD, so the generalization to other audio-video forgery datasets remains to be further validated.

## Acknowledgments

We thank the DDL-GAV challenge organizers for providing the MVAD benchmark and evaluation platform. This work was supported by the National Natural Science Foundation of China (No. 62472385), NSFC (Joint Fund) Key Program (No. U25A20440), Pioneer and Leading Goose R&D Program of Zhejiang (No. 2024C01110), Fundamental Research Funds for the Provincial Universities of Zhejiang (No. FR2402ZD), Tongxiang Institute of Artificial General Intelligence (No. TAGI2-B-2024-0004) and Zhejiang Provincial High-Level Talent Special Support Program (2022R52048).

## Ethical Statement

This work uses only the public MVAD benchmark and is intended to support responsible research on AI-generated audio-video detection.

## References


[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*, pages 2278–2282, 2022.

[Bar-Tal *et al.*, 2024] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[Blattmann *et al.*, 2023] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

[Brooks *et al.*, 2024] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, February 2024.

[Cai *et al.*, 2023] Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization. *Computer Vision and Image Understanding*, 236:103818, 2023.

[Cai *et al.*, 2024] 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. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7414–7423, 2024.

[Chen *et al.*, 2023] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR, 2023.

[Chen *et al.*, 2024] Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Zhao Lv, and Cunhang Fan. RawBMamba: End-to-end bidirectional state space model for audio deepfake detection. In *Proceedings of Interspeech 2024*, pages 2720–2724, 2024.

[Chen *et al.*, 2026] Haoran Chen, Daiqi Gao, Daiqi Li, Xinyuan Li, Yongwei Wang, Lei Zhu, Rui Cai, and Jianfeng Dong. Learning to falsify: Hierarchical multiple instance learning for high resolution ai-generated image detection. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026.

[Copet *et al.*, 2023] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[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. In *Interspeech 2024*, pages 547–551, 2024.

[Haliassos *et al.*, 2021] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021.

[Hu *et al.*, 2026] Mengxue Hu, Yunfeng Diao, Changtao Miao, Tairui Ge, Taize Ge, Zhiqing Guo, Jianshu Li, Zhe Li, Zhongjie Ba, and Joey Tianyi Zhou. MVAD: A benchmark dataset for multimodal ai-generated video-audio detection, 2026.

[Jung *et al.*, 2022] 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. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6367–6371, 2022.

[Katamneni and Rattani, 2024] Vinaya Sree Katamneni and Ajita Rattani. Contextual cross-modal attention for audio-visual deepfake detection and localization. In *IEEE International Joint Conference on Biometrics*, pages 1–11, 2024.

[Khalid *et al.*, 2021] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[Li *et al.*, 2020] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[Liu *et al.*, 2021] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021.

[Liu *et al.*, 2023] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th*

*International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR, 2023.

[Liu *et al.*, 2024] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

[Nguyen *et al.*, 2024] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024.

[Nguyen *et al.*, 2025] Dat Nguyen, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2025.

[Oorloff *et al.*, 2024] Trevine Oorloff, Surya Koppisetti, Nicolo Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27092–27102, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.

[Ouajdi *et al.*, 2024] Hafsa Ouajdi, Oussama Hadder, Modan Tailleur, Mathieu Lagrange, and Laurie M. Heller. Detection of deepfake environmental audio. In *Proceedings of the 32nd European Signal Processing Conference*, pages 196–200, 2024.

[Polyak *et al.*, 2024] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models, 2024.

[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, 2025.

[Tak *et al.*, 2022] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*, pages 112–119, 2022.

[Wang *et al.*, 2023a] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023.

[Wang *et al.*, 2023b] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. AltFreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 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.

[Yang *et al.*, 2025] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Zhang *et al.*, 2023] Yibo Zhang, Weiguo Lin, and Junfeng Xu. Joint audio-visual attention with contrastive learning for more general deepfake detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(5):1–23, 2023.

[Zhang *et al.*, 2026] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026.

[Zheng *et al.*, 2021] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15044–15054, 2021.

[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*, 2025.

[Zhou and Lim, 2021] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.