Less is More: Modality-Decoupling for General AIGC Audio-Video Detection

Jielun Peng[1], **Yabin Wang**[1], **Yaqi Li**[1], **Jincheng Liu**[1], **Xiaopeng Hong**[1]*
and **Athanasios V. Vasilakos**[2]
[1]Harbin Institute of Technology [2]University of Agder
{25s003052, 25s103223, 25b903114}@stu.hit.edu.cn, wang-yabin@outlook.com,
hongxiaopeng@ieee.org, th.vasilakos@gmail.com

Abstract

Generative AI has rapidly expanded audio-visual forgery beyond human-centric deepfakes into general scenes. Existing AIGC detection methods assume audio-visual content correspondence, identifying forgeries by spotting cross-modal inconsistencies. However, we empirically find that this assumption does not consistently hold in general scenarios. We argue that, for general audio-visual AIGC detection, decision-level fusion is a more robust alternative to feature-level fusion. Therefore, we propose DAV-Det, a decoupled audio-visual AIGC detection system that independently models forensic evidence from each modality. The visual detector leverages multi-granularity representations at global, patch, and segment levels to capture spatial forgery cues, while the audio detector exploits both temporal and spectral irregularities via a gated temporal-spectral dual-branch architecture to model acoustic artifacts. Our method ranks 1st in the General AIGC Audio-Video Detection Challenge of the IJCAI-ECAI 2026 DDL 2.0 Workshop, with a final score of 0.8460. Code is available at *https://github.com/tuffy-studio/DAV-Det*.

1 Introduction

The rapid advancement of Artificial Intelligence Generated Content (AIGC) technologies has revolutionized the creation of audio-visual content. While early AIGC systems primarily focused on human-centric generation, recent advances in generative models have enabled the synthesis of realistic audio-visual content in diverse scenarios beyond human faces and speech [Wang *et al.*, 2024; Wang *et al.*, 2025a; Chen *et al.*, 2026]. This shift substantially broadens the scope of potential misuse, ranging from misinformation to forged evidence. Consequently, reliable detection of general AIGC content has become increasingly important.

To combat the growing threat of AI-generated forgeries, numerous detection methods have been proposed. Some approaches focus on a single modality, detecting artifacts only from either audio [Jung *et al.*, 2022; Chen *et al.*, 2024; Gu *et al.*, 2025] or visual modalities [Wang *et al.*, 2025b; Wang *et al.*, 2026b]. Such methods fail to fully exploit the comprehensive information available across modalities. Building upon uni-modal limitations, multi-modal detection approaches [Yin *et al.*, 2024; Miao *et al.*, 2025; Oorloff *et al.*, 2024] have achieved remarkable success by jointly modeling audio and visual information. These methods are typically based on the assumption that, in human-centric scenarios, audio and visual signals exhibit content correspondence between speech phonemes and lip movements. Such correspondence is typically assessed via audio-visual matching measures, including frame-level alignment [Feng *et al.*, 2023; Smeu *et al.*, 2025] and feature similarity [Liang *et al.*, 2024; Peng *et al.*, 2026]. Manipulation of either modality may disrupt this correspondence, thereby providing discriminative cues for AIGC detection.

However, this assumption does not consistently hold in

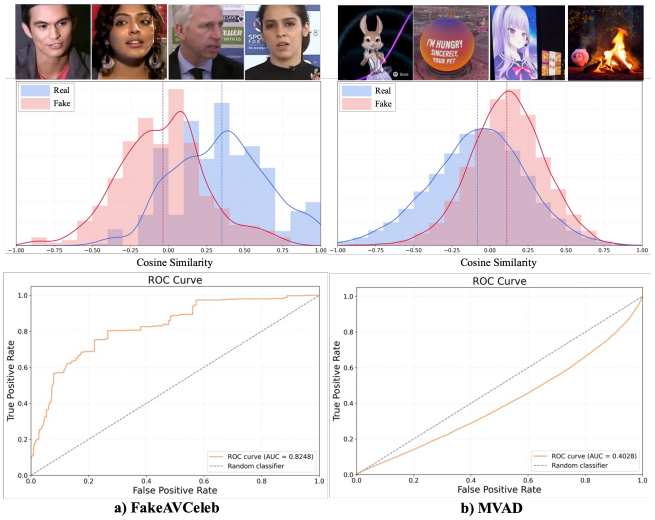


Figure 1: Distributions of cosine similarities between audio and video features extracted by PEAV. On FakeAVCeleb, real samples exhibit higher similarity than fake samples. In contrast, the audio-visual similarities significantly decrease in MVAD compared to human-centric datasets. Moreover, the similarities of real samples are generally even lower than those of fake ones in MVAD. Using this score alone for training-free detection yields an AUC below 0.5, suggesting that the traditional assumption of audio-visual correspondence is actively detrimental in general AIGC scenarios.

*Corresponding author.

general audio-visual scenarios. While human-centric content often exhibits content consistency between audio and visual streams, general videos tend to show weaker alignment in their underlying content. To validate our hypothesis, we use PEAV [Vyas *et al.*, 2026] to quantify audio-visual correspondence on two representative datasets: FakeAVCeleb [Khalid *et al.*, 2021] and MVAD [Hu *et al.*, 2025]. FakeAVCeleb is a widely used human-centric deepfake dataset, while MVAD contains general-scene AIGC content. As shown in Fig. 1, real samples in FakeAVCeleb exhibit higher audio-visual cosine similarity than fake samples. Following [Liang *et al.*, 2024; Smeu *et al.*, 2025], we rank samples based on the negative of cosine similarity and compute the area under the receiver operating characteristic curve (AUC), achieving a score of 82.48%. This result indicates that audio-visual correspondence serves as an effective cue for human-centric deepfake detection. However, on MVAD, fake samples even show higher similarity than real ones. Under the same evaluation protocol, where fake samples are expected to have lower audio-visual similarity, this leads to an AUC of only 40.28% (i.e., even worse than random guessing). As a result, the effectiveness of existing multi-modal detectors may be limited in general AIGC scenarios.

These findings motivate us to explore a simpler yet more robust paradigm: independently modeling modality-specific forensic cues and performing decision-level fusion. Accordingly, we propose DAV-Det, a Decoupled Audio-Visual AIGC Detection system that avoids explicit feature-level cross-modal interaction. Specifically, considering that AIGC cues may emerge at multiple spatial granularities, including global semantic anomalies, subtle local texture artifacts and regional inconsistencies, we develop a visual AIGC detector (see Fig. 2 (a)) that jointly exploits global, patch-, and segment-level representations to capture forgery cues at different levels. Complementarily, we design an audio AIGC detector (see Fig. 2 (b)) that exploits both temporal dependencies and spectral irregularities via a gated temporal-spectral dual-branch architecture, enabling fine-grained modeling of acoustic artifacts. During inference, the two detectors operate independently and their predictions are fused to enable both binary and four-class classification.

Our main contributions are summarized as follows:

- We empirically demonstrate that the assumption of audio-visual content correspondence, which underlies most existing multi-modal deepfake detectors, does not consistently hold in general AIGC scenarios.
- We propose DAV-Det, which independently models audio and visual forensic cues via multi-granularity visual representations and an audio temporal-spectral design, and performs decision-level fusion for audio-visual AIGC detection.
- Our method demonstrates competitive performance in the General AIGC Audio-Video Detection Challenge of the DDL 2.0 Workshop at IJCAI-ECAI 2026, ranking 1st with a final score of 0.8460.

2 Related Works

2.1 Uni-modal AIGC Detection

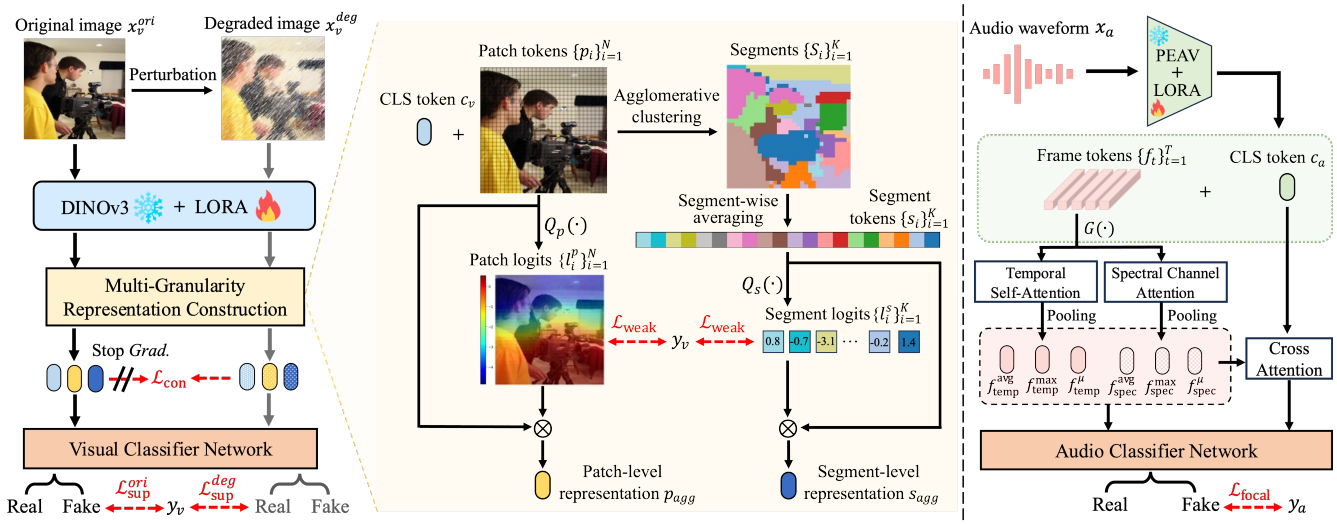
Visual-only methods. Methods leveraging visual artifacts focus on spatial and temporal inconsistencies, such as abnormal regions [Zhao *et al.*, 2021] and irregular movements [Haliassos *et al.*, 2021]. To improve generalization, prior works explore data augmentation with synthetic samples. Representative methods include artifact simulation, e.g., blending boundaries [Shiohara and Yamasaki, 2022; Chen *et al.*, 2022], and latent-space augmentation to enhance decision boundary [Yan *et al.*, 2024a; Dhakal *et al.*, 2026]. Another line of work is leveraging the pre-trained models to alleviate overfitting. For example, Effort [Yan *et al.*, 2024b] decomposes the feature space into orthogonal subspaces, preserving pre-trained knowledge while learning forgery cues. Query-based token refinement is proposed in [Wang *et al.*, 2026a] to mitigate CLS token bias inherited from pre-training and enhance the modeling of localized forgery cues.

Audio-only methods. Early methods [Wu *et al.*, 2017] rely on hand-crafted features (e.g., MFCC, LFCC) to capture synthesis artifacts, but generalize poorly to diverse channels and unseen generators. With the rise of deep learning, audio spoofing detectors have shifted toward end-to-end neural architectures. A representative baseline is AASIST [Jung *et al.*, 2022], which operates directly on raw waveforms using a Sinc-based convolutional front-end and residual blocks, followed by spectral–temporal attention for classification. To further improve robustness, recent approaches [Tak *et al.*, 2022; Lee *et al.*, 2025; Xie *et al.*, 2026] incorporate self-supervised learning (SSL) representations [Baevski *et al.*, 2020; Conneau *et al.*, 2020; Shi *et al.*, 2022]. By leveraging large-scale pretraining, these methods provide more transferable acoustic features and have become the dominant paradigm in modern audio AIGC detection.

However, despite achieving remarkable success in unimodal AIGC detection, these methods are inherently limited by their reliance on a single modality and fail to fully leverage the comprehensive cues available across modalities.

2.2 Audio-Visual AIGC Detection

Recent studies have increasingly explored leveraging both audio and visual modalities for more reliable AIGC detection. AVoid-DF [Yang *et al.*, 2023] detects multi-modal deepfakes by exploiting audio-visual inconsistency. AVGraph [Yin *et al.*, 2024] constructs heterogeneous audio-visual graphs to achieve fine-grained forgery classification. Several audio-visual methods also build upon pre-trained models. The authors of AVFF [Oorloff *et al.*, 2024] first pre-train the model to learn audio-visual correspondences in a self-supervised stage, followed by supervised deepfake classification. AVPrompt [Miao *et al.*, 2025] fine-tunes CLIP [Radford *et al.*, 2021] and Whisper [Radford *et al.*, 2023] for deepfake detection via prompt learning. HAVIC [Peng *et al.*, 2026] detects AIGC based on holistic audio-visual intrinsic coherence priors in the pre-trained models. In addition, unsupervised methods, including AVAD [Feng *et al.*, 2023], SpeechForensics [Liang *et al.*, 2024], and AVH-Align [Smeu *et al.*, 2025],



(a) Visual AIGC Detector **(b) Audio AIGC Detector**

Figure 2: **Overview of our proposed DAV-Det.** (a) The visual AIGC detector models global, patch-, and segment-level representations to capture forgery cues at multiple spatial granularities. (b) The audio AIGC detector characterizes temporal dynamics and spectral anomalies in the audio stream for comprehensive AIGC detection.

detect forgeries by measuring the matching degree between audio and video, without requiring ground-truth labels.

However, these methods are typically built upon the assumption of strong audio-visual correspondence in authentic content. In general videos, audio may be weakly related, or even unrelated to the visual content, making audio-visual inconsistency a natural characteristic rather than a sign of manipulation. Consequently, the effectiveness of existing multimodal detectors may be limited in general AIGC scenarios.

3 Methodology

3.1 Preprocessing

We first split the MVAD dataset [Hu *et al.*, 2025] into an audio dataset $\mathcal{D}_a = \{(x_i^a, y_i^a)\}$ and an image dataset $\mathcal{D}_v = \{(x_i^v, y_i^v)\}$. For $\mathcal{D}_a$, the waveform x_i^a is directly extracted from each sample, and $y_i^a \in \{0, 1\}$ indicates whether the audio is real (0) or fake (1). For $\mathcal{D}_v$, to balance classes, we uniformly sample 8 images from real videos and 16 images from fake videos to construct the dataset. All sampled images inherit the corresponding video-level label.

3.2 Visual AIGC Detector

Feature Extraction. Given an input image, we employ DINOv3 [Siméoni *et al.*, 2025] as the backbone to extract visual features, yielding a CLS token $c \in \mathbb{R}^D$ and a set of patch tokens $\{p_i\}_{i=1}^N \in \mathbb{R}^{N \times D}$, where N is the number of patches and D is the feature dimension. The CLS token captures holistic information of the image, while the patch tokens preserve fine-grained spatial local details.

Multi-Granularity Representation Construction. AIGC visual cues may emerge at multiple spatial granularities, from subtle local texture artifacts and regional inconsistencies to global semantic anomalies. To capture such multi-scale cues, we construct a multi-granularity representation consisting of global, patch-level, and segment-level representations.

Since the CLS token c_v serves as a global summary by aggregating information from all image patches through self-attention [Dosovitskiy *et al.*, 2020], we directly adopt it as the global representation.

For patch-level representation, we first employ a two-layer MLP as a patch authenticity scoring network $Q_p(\cdot)$ to estimate forgery logits $\{l_i^p\}_{i=1}^N$ for each patch token. The logits are normalized via a softmax function with temperature τ:

$$w_i^p = \frac{\exp(l_i^p/\tau)}{\sum_{j=1}^N \exp(l_j^p/\tau)}, \tag{1}$$

and used as aggregation weights to compute a weighted sum of patch tokens, producing the patch-level representation:

$$p_{agg} = \sum_{i=1}^N w_i^p p_i. \tag{2}$$

For segment-level representation, inspired by [Cuttano *et al.*, 2026], we partition the patch tokens into K disjoint spatial segments $\mathcal{S} = \{S_1, \ldots, S_K\}$ via agglomerative clustering [Müllner, 2011] based on feature cosine similarity. The clustering procedure progressively merges similar neighboring patches in a bottom-up manner, yielding

$$\bigcup_{k=1}^K S_k = \{p_i\}_{i=1}^N, \qquad S_i \cap S_j = \emptyset \quad \forall i \neq j. \tag{3}$$

For each segment S_k, a segment token s_k is obtained by average pooling the patch tokens within the segment:

$$s_k = \frac{1}{|S_k|} \sum_{p_i \in S_k} p_i. \tag{4}$$

A two-layer MLP is then employed as a segment authenticity scoring network $Q_s(\cdot)$ to estimate forgery logits $\{l_k^s\}_{k=1}^K$

for each segment token. The logits are normalized using a softmax function with temperature τ:

$$w_k^s = \frac{\exp(l_k^s/\tau)}{\sum_{j=1}^{K} \exp(l_j^s/\tau)}. \tag{5}$$

Similarly, the normalized logits are then used as aggregation weights to compute the segment-level representation:

$$s_{agg} = \sum_{k=1}^{K} w_k^s s_k. \tag{6}$$

Weak Supervision via Multiple Instance Learning. Since only image-level labels are available, the patch and segment authenticity scoring networks, $Q_p(\cdot)$ and $Q_s(\cdot)$, lack explicit supervision at the patch and segment levels. We therefore adopt a Multiple Instance Learning strategy to provide weak supervision for both networks. Following the assumption that manipulated images typically contain a portion of highly suspicious patches or segments, the top-k forgery logits are aggregated to obtain image-level predictions:

$$l_{\text{topk}}^p = \frac{1}{|\mathcal{T}_p|} \sum_{i \in \mathcal{T}_p} l_i^p, \quad l_{\text{topk}}^s = \frac{1}{|\mathcal{T}_s|} \sum_{i \in \mathcal{T}_s} l_i^s, \tag{7}$$

where $\mathcal{T}_*$ denotes the set of the top-k instances ranked by logits. The corresponding Multiple Instance Learning loss is

$$\mathcal{L}_{\text{MIL}} = \mathcal{L}_{\text{focal}}\big(y, \sigma(l_{\text{topk}}^p)\big) + \mathcal{L}_{\text{focal}}\big(y, \sigma(l_{\text{topk}}^s)\big), \tag{8}$$

where $y \in \{0, 1\}$ denotes the image-level ground-truth label, $\mathcal{L}_{\text{focal}}$ denotes the focal loss [Lin *et al.*, 2017], and $\sigma(\cdot)$ denotes the sigmoid function.

To further encourage the model to distinguish suspicious instances from the remaining ones, we introduce a margin loss. Specifically, we compute the forgery logits gap between the top-k instances and the remaining instances:

$$d_p = \frac{1}{|\mathcal{T}_p|} \sum_{i \in \mathcal{T}_p} l_i^p - \frac{1}{N - |\mathcal{T}_p|} \sum_{i \notin \mathcal{T}_p} l_i^p. \tag{9}$$

$$d_s = \frac{1}{|\mathcal{T}_s|} \sum_{i \in \mathcal{T}_s} l_i^s - \frac{1}{K - |\mathcal{T}_s|} \sum_{i \notin \mathcal{T}_s} l_i^s. \tag{10}$$

The margin loss is formulated as

$$\mathcal{L}_{\text{margin}} = \max(m - d_p, 0) + \max(m - d_s, 0), \tag{11}$$

where m denotes a predefined margin.

The overall weak supervision loss is defined as

$$\mathcal{L}_{\text{weak}} = \mathcal{L}_{\text{MIL}} + \mathcal{L}_{\text{margin}}. \tag{12}$$

Classifier Network. We employ a three-layer MLP that serves as the main classifier, taking the concatenated multi-granularity representation c_v, p_{agg}, and s_{agg} as input to classify whether an image is real or fake. To prevent the classifier from relying excessively on a particular representation level, we introduce three auxiliary classifiers that operate on c_v, p_{agg}, and s_{agg}, respectively. All classifiers are trained using the focal loss [Lin *et al.*, 2017] with the label y_v as supervision, while only the main classifier is used during inference. The supervised loss is defined as the sum of the focal losses from the main classifier and the three auxiliary classifiers:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{segment}}. \tag{13}$$

Degraded-Original Consistency Learning. To improve the robustness of the detector, we introduce diverse image perturbations as data augmentation, including noise, blur, geometric transformations, and color jittering, to simulate real-world image degradations. Based on these augmented views, to encourage the model to learn degradation-robust representation, we further introduce a Degraded-Original Consistency Learning (DOCL) strategy. For each original input image x_v^{ori}, we generate a degraded counterpart x_v^{deg} using the perturbation pipeline. x_v^{ori} and x_v^{deg} are separately fed into the backbone. A consistency loss $\mathcal{L}_{\text{con}}$ is employed to enforce multi-granularity representation consistency between the original and degraded images:

$$\mathcal{L}_{\text{con}} = \sum_{z \in \{c_v, p_{agg}, s_{agg}\}} \big(1 - \text{sim}\big(\text{sg}[z^{ori}], z^{deg}\big)\big), \tag{14}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, and $\text{sg}[\cdot]$ represents the stop-gradient operation, meaning that gradients from $\mathcal{L}_{\text{con}}$ are only propagated through the degraded representations, while the representations of the original image are treated as fixed targets.

Finally, the overall visual loss function is defined as:

$$\mathcal{L}_v = \mathcal{L}_{\text{sup}}^{ori} + \mathcal{L}_{\text{weak}}^{ori} + \mathcal{L}_{\text{sup}}^{deg} + \mathcal{L}_{\text{weak}}^{deg} + 0.05 * \mathcal{L}_{\text{con}}. \tag{15}$$

3.3 Audio AIGC Detector

Feature Extraction. We leverage the audio encoder of PEAV [Vyas *et al.*, 2026] as the audio backbone. The input audio waveform is fed into the backbone to extract audio features. The output comprises a CLS token $c_a \in \mathbb{R}^D$ and a sequence of frame tokens $\{f_t\}_{t=1}^T \in \mathbb{R}^{T \times D}$, where T denotes the number of audio frames. The CLS token captures holistic acoustic properties, while the frame tokens preserve fine-grained temporal-spectral details.

Temporal-Spectral Representation Construction. AIGC audio artifacts manifest in both temporal dynamics (e.g., prosody inconsistency and unnatural transitions) and spectral characteristics (e.g., high-frequency artifacts). To capture these multi-aspect cues, inspired by [Jung *et al.*, 2022], we design a temporal-spectral dual-branch architecture.

To adaptively modulate discriminative temporal-spectral features, we compute a feature-wise gating map via a gating network $G(\cdot)$, implemented as a two-layer MLP:

$$W = \sigma(G(\{f_t\}_{t=1}^T)) \in (0, 1)^{T \times D}, \tag{16}$$

where σ denotes the sigmoid function. The gating map is then used to modulate the audio frame token sequence:

$$\{\tilde{f}_t\}_{t=1}^T = \{f_t\}_{t=1}^T \odot W. \tag{17}$$

The temporal branch first refines the modulated frame sequence $\{\tilde{f}_t\}_{t=1}^T$ by modeling temporal dependencies across frames using self-attention [Vaswani *et al.*, 2017], yielding a temporally enhanced feature sequence F_{temp}. Based on F_{temp},

Table 1: Perturbation methods used in the Degraded-Original Consistency Learning.

Category	Methods
Geometric Transformations	Horizontal Flip, Image Rotation, Image Translation, Image Resizing
Noise	Gaussian Noise, Impulse Noise, Speckle Noise, Poisson Noise
Blur	Gaussian Blur, Lens Defocus Blur, Mean Blur
Color Distortion	Saturation Change, Grayscale, Quantization
Photometric Transformations	Brighten, Darken, Contrast Change
Compression	JPEG Compression

we construct complementary temporal representations that summarize different aspects of the temporal dynamics, including an average token $f_{\text{temp}}^{\text{avg}}$ via temporal average pooling, a discriminative token $f_{\text{temp}}^{\max}$ via max pooling, and an adaptive token f_{temp}^{μ} via learnable query-based pooling.

The spectral branch operates on the same modulated features and employs a channel attention [Hu *et al.*, 2018] mechanism to emphasize spectral-related patterns. This process yields a spectrally enhanced feature sequence F_{spec}. Based on F_{spec}, we derive complementary spectral representations via different pooling strategies in the spectral domain, including an average token $f_{\text{spec}}^{\text{avg}}$, a discriminative token $f_{\text{spec}}^{\max}$, and a learnable query-based token f_{spec}^{μ}.

To integrate global and local cues, the CLS token c_a attends to the above representations via cross-attention [Vaswani *et al.*, 2017]. This process allows the model to selectively aggregate informative signals from both temporal and spectral perspectives, producing a unified audio representation m.

Classifier Network. The final representation $\mathbf{r}$ is constructed by concatenating the unified audio representation with temporal and spectral representations:

$$\mathbf{r} = \text{Concat}[m, f_{\text{temp}}^{\text{avg}}, f_{\text{temp}}^{\max}, f_{\text{temp}}^{\mu}, f_{\text{spec}}^{\text{avg}}, f_{\text{spec}}^{\max}, f_{\text{spec}}^{\mu}]. \quad (18)$$

We employ a three-layer MLP as the audio classifier, which takes $\mathbf{r}$ as input and predicts whether the input audio is real or fake. The classifier is trained using Focal Loss [Lin *et al.*, 2017] with audio label y_a as supervision.

3.4 Decision-Level Fusion for Inference

During inference, the audio and visual detectors independently produce fake probabilities for each input. For the visual stream, we uniformly sample 16 frames from each video clip. The visual detector produces frame-level logits, which are then averaged to obtain the final video-level logit $\bar{l}_v$. A sigmoid function is applied to obtain the fake probability:

$$p_v^{Fake} = \sigma(\bar{l}_v). \quad (19)$$

For the audio stream, we adopt a sliding window strategy with a window size and stride of 3 seconds. The audio waveform is segmented into consecutive windows, and the audio detector produces window-level logits. These logits are then averaged to obtain the final audio-level logit $\bar{l}_a$. A sigmoid function is applied to obtain the fake probability:

$$p_a^{Fake} = \sigma(\bar{l}_a). \quad (20)$$

The corresponding real probabilities are defined as:

$$p_a^{Real} = 1 - p_a^{Fake}, \quad p_v^{Real} = 1 - p_v^{Fake}. \quad (21)$$

For binary classification, we adopt a max-based fusion strategy. The final fake probability is defined as:

$$p^{Fake} = \max(p_a^{Fake}, p_v^{Fake}), \quad (22)$$

and the real probability is computed as:

$$p^{Real} = 1 - p^{Fake}. \quad (23)$$

This design assumes that a sample is fake if either the audio or visual modality exhibits a high fake probability.

For four-class classification, we define a joint audio-visual probability distribution under an independence assumption:

$$\begin{aligned} p_{RR} &= p_v^{Real} \cdot p_a^{Real}, \\ p_{FF} &= p_v^{Fake} \cdot p_a^{Fake}, \\ p_{FR} &= p_v^{Fake} \cdot p_a^{Real}, \\ p_{RF} &= p_v^{Real} \cdot p_a^{Fake}, \end{aligned} \quad (24)$$

where p_{RR} denotes the case where both modalities are real, and p_{FF} indicates that both are fake. p_{FR} and p_{RF} correspond to the video being fake while the audio is real, and the video being real while the audio is fake, respectively. The final prediction is obtained by selecting the class with the highest joint probability.

4 Experiments

4.1 Experimental Setup

Dataset and metrics. We evaluate our method on the General AIGC Audio–Video Detection benchmark (DDL-GAV), which is built upon the MVAD dataset [Hu *et al.*, 2025]. DDL-GAV is a large-scale multimodal benchmark containing over 200,000 video-audio samples covering both realistic and anime visual domains, four content categories (humans, animals, objects, and scenes), and more than 23 state-of-the-art AIGC generation methods. Based on which modalities are generated by AI, the samples can be categorized into four types: real video with fake audio (RF), fake video with real audio (FR), fake video with fake audio (FF), and real video with real audio (RR, i.e., the authentic samples).

The benchmark includes two hierarchical tasks: binary forgery detection and four-class modality classification. For binary detection, samples are classified as either real or fake. For four-class classification, each sample belongs to one of the following categories: RF, FR, FF, and RR. Following the official evaluation protocol, model performance is assessed using Area Under the ROC Curve (AUC), Accuracy (ACC),

Table 2: Results of the AIGC Audio-Video Detection Challenge (Top-5 teams). Best result is in **bold**, and second best is underlined.

Rank	Team	Final Score	Binary Score	Four-Class Score	Binary AUC	Binary ACC	Four-Class F1	Four-Class AP
1	**HIT VIRLAB (Ours)**	**0.8460**	0.9379	0.7847	**0.9617**	0.9190	**0.6873**	**0.7274**
2	ZJSU	0.8438	0.9395	0.7800	0.9490	**0.9217**	0.6847	0.7149
3	VeriLens	0.8426	0.9288	**0.7851**	0.9288	0.9106	0.6856	0.7096
4	AIGVDete	0.8406	**0.9396**	0.7746	0.9480	0.9187	0.6726	0.7191
5	MVADetection Team	0.8349	0.9278	0.7729	0.9447	0.9000	0.6530	0.7167

Average Precision (AP), and F1-score. For both binary and four-class classification, the official score is computed as:

$$\text{Score} = 0.2\times\text{AUC}+0.3\times\text{ACC}+0.3\times\text{AP}+0.2\times\text{F1}. \quad (25)$$

The final score is computed as a weighted sum of the binary detection score and the four-class classification score, with weights of 0.4 and 0.6, respectively.

Implementation details. For the visual AIGC detector, we employ the DINOv3 ViT-L/16 [Siméoni *et al.*, 2025] with feature dimension $D = 1024$ as the backbone, and apply LoRA adaptation [Hu *et al.*, 2022] (rank 32, alpha 16) for parameter-efficient fine-tuning. We train the visual AIGC detector using the AdamW optimizer with a weight decay of 5e-2 and a learning rate of 1e-4, decayed using a cosine schedule. We train for 20 epochs with a linear warmup for 2 epochs using eight NVIDIA L20 GPUs with a total batch size of 112 and a gradient accumulation interval of 16. The softmax temperature parameter τ is fixed at 0.07. The predefined margin used in the margin loss is fixed at 0.6. The top-k ratios are set to 0.05 for patches and 0.1 for segments. For agglomerative clustering, we use average linkage with a cosine similarity threshold of 0.9. Data augmentations used for Degraded-Original Consistency Learning are shown in Table 1. For the audio AIGC detector, we employ the audio encoder of PEAV-base [Vyas *et al.*, 2026] with feature dimension $D = 1024$ as the backbone, and apply LoRA adaptation [Hu *et al.*, 2022] (rank 32, alpha 64) for parameter-efficient fine-tuning. We train the audio AIGC detector using the AdamW optimizer with a weight decay of 5e-2 and a learning rate of 1e-4, decayed using a cosine schedule. We train for 20 epochs with a linear warmup for 2 epochs using four NVIDIA L20 GPUs with a total batch size of 512.

4.2 Competition Results

Table 2 reports the top-5 results on the General AIGC Audio-Video Detection Challenge leaderboard[1]. Our team, HIT VIRLAB, achieves the best overall performance with a final score of 0.8460, outperforming all competing teams.

Beyond the overall ranking, our method demonstrates consistently strong performance across both binary and four-class classification tasks. In binary classification, we achieve the highest AUC of 0.9617, significantly surpassing the second-best result (+1.27%), while maintaining competitive accuracy. For the more challenging four-class classification, our method attains the best F1 score of 0.6873 and AP score of 0.7274, indicating superior capability in fine-grained modality-aware discrimination.

[1]https://www.codabench.org/competitions/15769/#/results-tab

It is worth noting that AIGVDete achieves the highest binary accuracy, while VeriLens obtains slightly better performance in terms of four-class F1 score. Nevertheless, our approach maintains a more balanced performance across all evaluation metrics, leading to the highest overall ranking on the leaderboard.

4.3 Comparison with AV Deepfake Detectors

We further compare our DAV-Det with recent audio-visual methods on the FakeAVCeleb dataset [Khalid *et al.*, 2021]. For fair comparisons, we follow the same data preprocessing and split protocol in [Yang *et al.*, 2023; Oorloff *et al.*, 2024; Datta *et al.*, 2025; Nie *et al.*, 2026], using 70% of the FakeAVCeleb samples for training and validation, and the remaining 30% for testing. Facial regions are cropped from video frames using FaceX-Zoo [Wang *et al.*, 2021] to mitigate background interference. We report binary classification ACC and AUC as evaluation metrics, consistent with prior works for fair comparisons. As shown in Table 3, DAV-Det achieves superior performance compared with existing audio-visual deepfake detectors, reaching 0.998 ACC and 0.999 AUC. Although FakeAVCeleb mainly contains human-centric deepfakes rather than general AIGC-generated videos, these results demonstrate the strong detection capability of DAV-Det on conventional human-centric audio-visual deepfake scenarios, benefiting from its effective modality-specific detectors and modality decoupling strategy.

Table 3: Comparison with recent audio-visual methods on FakeAVCeleb. Best result is in **bold**, and second best is underlined.

Method	ACC	AUC
VFD [Cheng *et al.*, 2023]	0.815	0.861
AVoiD-DF [Yang *et al.*, 2023]	0.837	0.892
MCL [Liu *et al.*, 2023]	0.860	0.896
MRDF-CE [Zou *et al.*, 2024]	0.941	0.924
AVFF [Oorloff *et al.*, 2024]	0.986	0.991
PIA [Datta *et al.*, 2025]	0.987	0.998
FoVB [Nie *et al.*, 2026]	0.985	0.997
DAV-Det (Ours)	**0.998**	**0.999**

4.4 Ablation Study

In this section, we conduct ablation studies to evaluate the effectiveness of each component in DAV-Det. As the ground-truth labels of the official DDL-GAV test set are unavailable, we first randomly sample a subset from the original training

set and then divide it into training and validation subsets with a ratio of 9:1. The model is trained on the training subset, and the ablation studies are conducted on the validation subset. For evaluating the audio and visual detectors separately, a video is considered fake only when the corresponding modality is manipulated, i.e., audio manipulation for the audio detector and visual manipulation for the visual detector.

Ablation on Visual Detector. We first analyze the effectiveness of the multi-granularity design in the visual detector. The baseline only uses the global branch for classification. We then introduce patch-level and segment-level branches to enrich multi-granularity representations. As shown in Tab. 4, introducing multi-granularity branches brings progressive improvements over the global-only baseline, with the full model achieving the best results, demonstrating the effectiveness of multi-granularity representation for deepfake detection.

Table 4: Effectiveness of the multi-granularity design.

Global	Patch	Segment	ACC	AP	AUC	F1
✓	✗	✗	0.9742	0.9695	0.9895	0.9181
✓	✓	✗	0.9757	0.9706	0.9916	0.9228
✓	✓	✓	**0.9836**	**0.9879**	**0.9969**	**0.9482**

In addition, we study the contribution of the weak supervision loss, the auxiliary classification head, and the DOCL strategy. As shown in Tab. 5, removing the weakly-supervised loss leads to a performance drop, indicating that it provides effective weak supervision for the patch and segment authenticity scoring networks. Similarly, removing the auxiliary classifiers degrades performance, suggesting their effectiveness in facilitating multi-granularity representations for capturing deepfake cues. Removing the DOCL strategy also results in performance degradation, demonstrating its effectiveness in improving robustness to input perturbations and maintaining multi-granularity consistency across degraded and original views of the same image.

Table 5: Ablation study on individual strategies in visual detector.

Method	ACC	AP	AUC	F1
w/o weak supervision loss	0.9764	0.9773	0.9920	0.9252
w/o auxiliary classifiers	0.9803	0.9856	0.9951	0.9437
w/o DOCL strategy	0.9750	0.9749	0.9925	0.9210
Ours	**0.9836**	**0.9879**	**0.9969**	**0.9482**

Ablation on Audio Detector. We further investigate the effectiveness of the proposed temporal-spectral dual-branch design in the audio detector. As shown in Tab. 6, removing both branches (i.e., only using the CLS token for classification) causes a noticeable performance degradation, indicating the importance of explicitly modeling temporal-spectral audio representations for audio AIGC detection. Introducing the temporal branch improves the detection performance over the baseline, demonstrating that temporal dynamics provide valuable forgery-related cues. Similarly, incorporating the spectral branch further enhances the detection capability, indicating that spectral information also contains discriminative cues for audio AIGC detection. Finally, combining both temporal and spectral branches achieves the best performance. These results verify that temporal and spectral information provide complementary evidence, and their joint modeling enables more effective audio AIGC detection.

Table 6: Effectiveness of the temporal-spectral design.

Temporal	Spectral	ACC	AP	AUC	F1
✗	✗	0.9700	0.9759	0.9914	0.9302
✓	✗	0.9736	0.9783	0.9917	0.9387
✗	✓	0.9724	0.9773	0.9904	0.9358
✓	✓	**0.9791**	**0.9829**	**0.9942**	**0.9435**

5 Conclusion, Limitations, and Future Work

In this paper, we study general AIGC audio-video detection and empirically demonstrate that the assumption of audio-visual content correspondence, which underlies most existing multi-modal deepfake detectors, does not consistently hold in general AIGC scenarios. To address this, we propose DAV-Det, a decoupled audio-visual detection system that models each modality independently and performs decision-level fusion. The visual detector leverages multi-granularity representations at global, patch, and segment levels to capture forgery cues at different spatial levels, while the audio detector exploits both temporal and spectral irregularities via a gated temporal-spectral dual-branch architecture to model acoustic artifacts. Our method achieves first place in the General AIGC Audio-Video Detection Challenge at the IJCAI-ECAI 2026 DDL 2.0 Workshop, with a final score of 0.8460.

Limitations. Despite the effectiveness of DAV-Det, several limitations remain. First, the visual detector does not explicitly model fine-grained temporal dependencies, such as motion inconsistencies across frames. Second, the decision-level fusion strategy relies on heuristic rules, including max-based fusion for binary classification and independence assumptions for four-class prediction. These choices may not fully leverage the potential of adaptive decision-level fusion.

Future Work. Future work will explore temporal visual modeling to capture dynamic forgery traces across frames and develop more principled audio-visual decision-level fusion strategies, such as adaptive weighting and learnable calibration, to further improve audio-visual AIGC detection.

Acknowledgements

This work is funded by the National Key R&D Program of China (No. 2025YFC3811300), the National Natural Science Foundation of China (Grant No. 62376070 and 62076195), the Fundamental Research Funds for the Central Universities (AUGA5710028726), and the China Postdoctoral Science Foundation (No. 2026M794773).

References

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A

framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[Chen *et al.*, 2022] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022.

[Chen *et al.*, 2024] Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Lv Zhao, and Cunhang Fan. Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. *arXiv preprint arXiv:2406.06086*, 2024.

[Chen *et al.*, 2026] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *Science China Information Sciences*, 69(6):162103, 2026.

[Cheng *et al.*, 2023] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–22, 2023.

[Conneau *et al.*, 2020] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[Cuttano *et al.*, 2026] Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21638–21648, 2026.

[Datta *et al.*, 2025] Soumyya Kanti Datta, Tanvi Ranga, Chengzhe Sun, and Siwei Lyu. Pia: Deepfake detection using phoneme-temporal and identity-dynamic analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1596–1606, 2025.

[Dhakal *et al.*, 2026] Aayush Dhakal, Subash Khanal, Srikumar Sastry, Jacob Arndt, Philipe Dias, Dalton Lunga, and Nathan Jacobs. Simlbr: Learning to detect fake images by learning to detect real images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35472–35482, 2026.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Feng *et al.*, 2023] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

[Gu *et al.*, 2025] Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11736–11745, 2025.

[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.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[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.

[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 (ICASSP)*, pages 6367–6371. IEEE, 2022.

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

[Lee *et al.*, 2025] Seungeun Lee, Sunmook Choi, Taein Kang, Sanghyeok Chung, Soyul Han, Jaejin Seo, Seoyoung Park, Eujin Kim, Seungsang Oh, and Il-Youp Kwak. iwax: interpretable wav2vec-aasist-xgboost framework for voice spoofing detection. *Scientific reports*, 15(1):40491, 2025.

[Liang *et al.*, 2024] Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Boquan Li, Feng Yu, Ning Zhang, Xiang Meng, and Weiqing Huang. Speechforensics: Audio-visual speech representation learning for face forgery detection. *Advances in Neural Information Processing Systems*, 37:86124–86144, 2024.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2023] Xiaolong Liu, Yang Yu, Xiaolong Li, and Yao Zhao. Mcl: multimodal contrastive learning for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2803–2813, 2023.

[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.

[Müllner, 2011] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

[Nie *et al.*, 2026] Fan Nie, Jiangqun Ni, Jian Zhang, Bin Zhang, Weizhe Zhang, and Bin Li. Towards generalizable deepfake detection via forgery-aware audio-visual adaptation: A variational bayesian approach. *IEEE Transactions on Information Forensics and Security*, 2026.

[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.*, 2026] Jielun Peng, Yabin Wang, Yaqi Li, Long Kong, and Xiaopeng Hong. Leave no stone unturned: Uncovering holistic audio-visual intrinsic coherence for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6655–6666, 2026.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[Shi *et al.*, 2022] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022.

[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.

[Smeu *et al.*, 2025] Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18815–18825, 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. *arXiv preprint arXiv:2202.12233*, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Vyas *et al.*, 2026] Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30172–30182, 2026.

[Wang *et al.*, 2021] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM international conference on Multimedia*, pages 3779–3782, 2021.

[Wang *et al.*, 2024] Yabin Wang, Zhiwu Huang, Zhiheng Ma, and Xiaopeng Hong. Linguistic profiling of deepfakes: An open database for next-generation deepfake detection. *arXiv preprint arXiv:2401.02335*, 2024.

[Wang *et al.*, 2025a] Yabin Wang, Xiaopeng Hong, and Zhiwu Huang. Benchmark dataset and framework for continual ai-generated image detection. *Journal of Image and Graphics*, 30(11):3438–3450, 2025.

[Wang *et al.*, 2025b] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Opensdi: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4291–4301, 2025.

[Wang *et al.*, 2026a] Changshuo Wang, Jiangming Wang, Ke-Yue Zhang, Taiping Yao, Shouhong Ding, Shunli Wang, Ran Yi, and Lizhuang Ma. Beyond [cls] token: Query-driven token-level forgery purification for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 42922–42931, 2026.

[Wang *et al.*, 2026b] Yabin Wang, Zhiwu Huang, Zhou Su, Adam Prugel-Bennett, and Xiaopeng Hong. Penny-wise and pound-foolish in ai-generated image detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[Wu *et al.*, 2017] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and

Hector Delgado. Asvspoof: The automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):588–604, 2017.

[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.

[Yan *et al.*, 2024a] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024.

[Yan *et al.*, 2024b] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

[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.

[Zhao *et al.*, 2021] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[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.