MVADETECTOR: Modality-Specific Manifold Contrastive Learning for General AIGC Audio-Video Detection

Yuning Zhang[1], **Mingyu Liao**[1], **Tingyu Liu**[1], **Xinghao Wang**[2]
[1]Individual Researcher, [2]HiDream
ynzhang2022@gmail.com, takina.chris@gmail.com, tingyuliu119@gmail.com, wxhsdhr@gmail.com

Abstract

This report presents MVADETECTOR, our submission to Track 2 of the IJCAI-ECAI 2026 AI Safety Workshop, for general audio-visual AI-generated content (AIGC) detection. MVADETECTOR adopts a modality-specific representation learning framework that decomposes audio-visual authenticity prediction into video-side and audio-side evidence modeling. Specifically, we combine a VideoMAE-base encoder for visual streams and a Wav2Vec2-base encoder for audio streams, followed by lightweight projection heads and a late-fusion classifier for four-way authenticity classification. Rather than relying on explicit lip-sync modeling or cross-modal attention, we regularize each modality independently using triplet contrastive objectives constructed from authenticity-preserving and structure-disrupting augmentations. This design is well aligned with the challenge setting, where each label corresponds to a combination of video and audio authenticity. To address class imbalance, we further employ class-balanced cross-entropy and a weak binary auxiliary objective. Our best checkpoint achieves a binary score of 0.9241, a four-class score of 0.7674, and a final score of 0.8301.

1 Introduction

Recent advances in video and audio generation have substantially improved the realism of synthetic media, posing new challenges for media provenance, multimedia forensics, and AI safety. Track 2 of the IJCAI-ECAI 2026 AI Safety Workshop targets general audio-visual AI-generated content (AIGC) detection, where systems are required to identify whether the visual and audio streams of a video are authentic or generated [IJCAI-ECAI 2026 AI Safety Workshop Organizers, 2026]. The corresponding DDL-GAV benchmark is built upon the MVAD dataset [Hu *et al.*, 2026], covering over 100,000 audio-visual samples across diverse visual styles, content categories, and generation pipelines, which makes it a challenging testbed for generalizable multimodal authenticity analysis.

The task involves both binary authenticity detection and fine-grained four-way classification. The binary setting distinguishes fully authentic samples from those containing at least one manipulated modality, while the four-way setting predicts one of four modality-wise authenticity combinations: **RR** (real video-real audio), **RF** (real video-fake audio), **FR** (fake video-real audio), and **FF** (fake video-fake audio). Since the final score combines binary and four-class performance, a competitive system must not only detect the presence of synthetic content, but also identify the manipulated modality.

A central challenge lies in the highly imbalanced label distribution. In particular, the **FR** class is significantly underrepresented compared with the dominant **RR** and **FF** classes. Under standard classification training, models tend to bias toward frequent classes and misclassify rare **FR** samples as **FF**, which is especially detrimental under macro-averaged four-class evaluation. This imbalance suggests that the detector should learn modality-discriminative evidence rather than relying solely on global class priors.

Prior audio-visual deepfake research has progressed from multimodal benchmark construction to joint and modality-aware forensic modeling. FakeAVCeleb introduced a dedicated audio-video multimodal deepfake benchmark [Khalid *et al.*, 2021], while LAV-DF [Cai *et al.*, 2022] and AV-Deepfake1M [Cai *et al.*, 2024] further extended the problem toward temporally localized and large-scale multimodal forgeries. Existing detection approaches exploit affective discrepancies between modalities [Mittal *et al.*, 2020], explicitly learn audio-visual correspondence [Oorloff *et al.*, 2024], or disentangle modality-invariant and modality-specific representations [Katamneni and Rattani, 2023]. Closely related formulations also model modality-wise authenticity through four audio-video authenticity combinations [Muppalla *et al.*, 2023] or dual-label prediction [Yu *et al.*, 2023]. In contrast, our framework regularizes visual and audio representations independently with modality-specific structure-preserving and structure-disrupting views, and performs only lightweight late fusion for the final four-way decision.

We propose MVADETECTOR, a compact and reproducible audio-visual AIGC detection framework tailored to the structure of the challenge. Our key observation is that the four labels can be factorized into video-side and audio-side authenticity factors. Accordingly, MVADETECTOR learns

modality-specific representations using a VideoMAE-based visual encoder [Tong *et al.*, 2022] and a Wav2Vec2-based audio encoder [Baevski *et al.*, 2020], regularizes each modality with contrastive objectives, and performs late fusion only at the classifier level. This design avoids heavy cross-modal interaction modules while preserving the ability to infer modality-wise authenticity.

Our contributions are summarized as follows:

- We present a *modality-factorized* audio-visual AIGC detector that combines VideoMAE and Wav2Vec2 representations through a lightweight late-fusion classifier.
- We introduce a *modality-specific contrastive regularization strategy* that exploits authenticity-related structural variations through modality-aware augmentation.
- We mitigate the imbalanced four-class setting using square-root inverse-frequency class weights, weighted sampling, and a weak binary auxiliary objective.

2 Method

.

2.1 Overview

Given an input audio-visual clip, MVADETECTOR first separates it into a visual stream and an audio stream. The visual stream is represented by uniformly sampled video frames, while the audio stream is represented by a fixed-length waveform. The two streams are encoded by a VideoMAE-base visual encoder [Tong *et al.*, 2022] and a Wav2Vec2-base audio encoder [Baevski *et al.*, 2020], respectively. Each encoder output is then mapped by a lightweight two-layer projection head into a 256-dimensional *modality-specific* embedding. Finally, the two embeddings are concatenated and fed into a two-layer MLP classifier to predict four-way authenticity logits.

Figure 1 illustrates the overall training pipeline. In addition to the original visual and audio inputs, MVADETECTOR constructs modality-specific augmented views for contrastive regularization. For each modality, a *manifold-preserving* augmentation produces an authenticity-preserving positive view, whereas a *manifold-disrupting* augmentation produces a structure-disrupted pseudo-negative view. The final learning objective combines four-class cross-entropy, video-side contrastive loss, audio-side contrastive loss, and a weak binary auxiliary loss.

2.2 Video Encoder

The visual branch adopts VideoMAE-base [Tong *et al.*, 2022] as the backbone encoder. For each clip, we uniformly sample 16 frames from the full video and resize each frame to 224×224. The resulting frame sequence is fed into VideoMAE [Tong *et al.*, 2022], which outputs a sequence of visual tokens. We use the CLS token as the clip-level visual representation. We also evaluated mean pooling over all tokens, but it did not improve the final leaderboard score.

$$z_v = \mathrm{Proj}_v\big(\mathrm{CLS}(\mathrm{VideoMAE}(x_v))\big). \tag{1}$$

Here, x_v denotes the input visual stream, Proj_v denotes the visual projection head, and z_v is the resulting 256-dimensional visual embedding.

2.3 Audio Encoder

The audio branch adopts Wav2Vec2-base [Baevski *et al.*, 2020] as the backbone encoder. The audio track is decoded at 16 kHz and converted into a fixed-length 4-second waveform. Longer waveforms are cropped during training, while shorter ones are zero-padded to the target length. The waveform is then fed into Wav2Vec2 [Baevski *et al.*, 2020] to extract frame-level audio representations.

We apply mean pooling over the temporal dimension of the last hidden state and map the pooled representation to a 256-dimensional audio embedding:

$$z_a = \mathrm{Proj}_a\big(\mathrm{MeanPool}(\mathrm{Wav2Vec2}(x_a))\big). \tag{2}$$

Here, x_a denotes the input audio stream, Proj_a denotes the audio projection head, and z_a is the resulting 256-dimensional audio embedding.

2.4 Fusion Classifier

The modality-specific embeddings are concatenated into a 512-dimensional joint representation and passed to a lightweight late-fusion classifier to produce four-way authenticity logits:

$$\ell = \mathrm{MLP}([z_v; z_a]). \tag{3}$$

The classifier is implemented as a two-layer MLP with a 256-dimensional hidden layer, GELU activation, and dropout regularization.

We adopt *late fusion* rather than explicit attention-based cross-modal interaction. This design is motivated by the factorized label structure of the task: each target class corresponds to a specific combination of visual authenticity and audio authenticity. Therefore, the visual and audio embeddings provide modality-wise evidence for the two axes of the four-way decision, while the final classifier integrates these cues for joint prediction.

2.5 Modality-Specific Manifold Contrastive Learning

Different from conventional contrastive learning that mainly aims to learn semantic invariance, our objective is to emphasize authenticity-related structures. In AIGC detection, some transformations only modify nuisance factors while preserving the underlying authenticity characteristics, whereas others intentionally disturb temporal or spectral regularities that are closely related to authenticity-related artifacts. Therefore, we construct two types of modality-aware views: authenticity-preserving views and structure-disrupting views.

A key challenge in audio-visual AIGC detection is to learn modality-wise evidence that is robust to benign variations but sensitive to authenticity-related structural artifacts. For the visual stream, such evidence often appears in temporal consistency, frame-level continuity, and local visual artifacts. For the audio stream, it may be reflected in prosodic continuity, spectral structure, and phase consistency. Motivated by this observation, we introduce *modality-specific manifold*

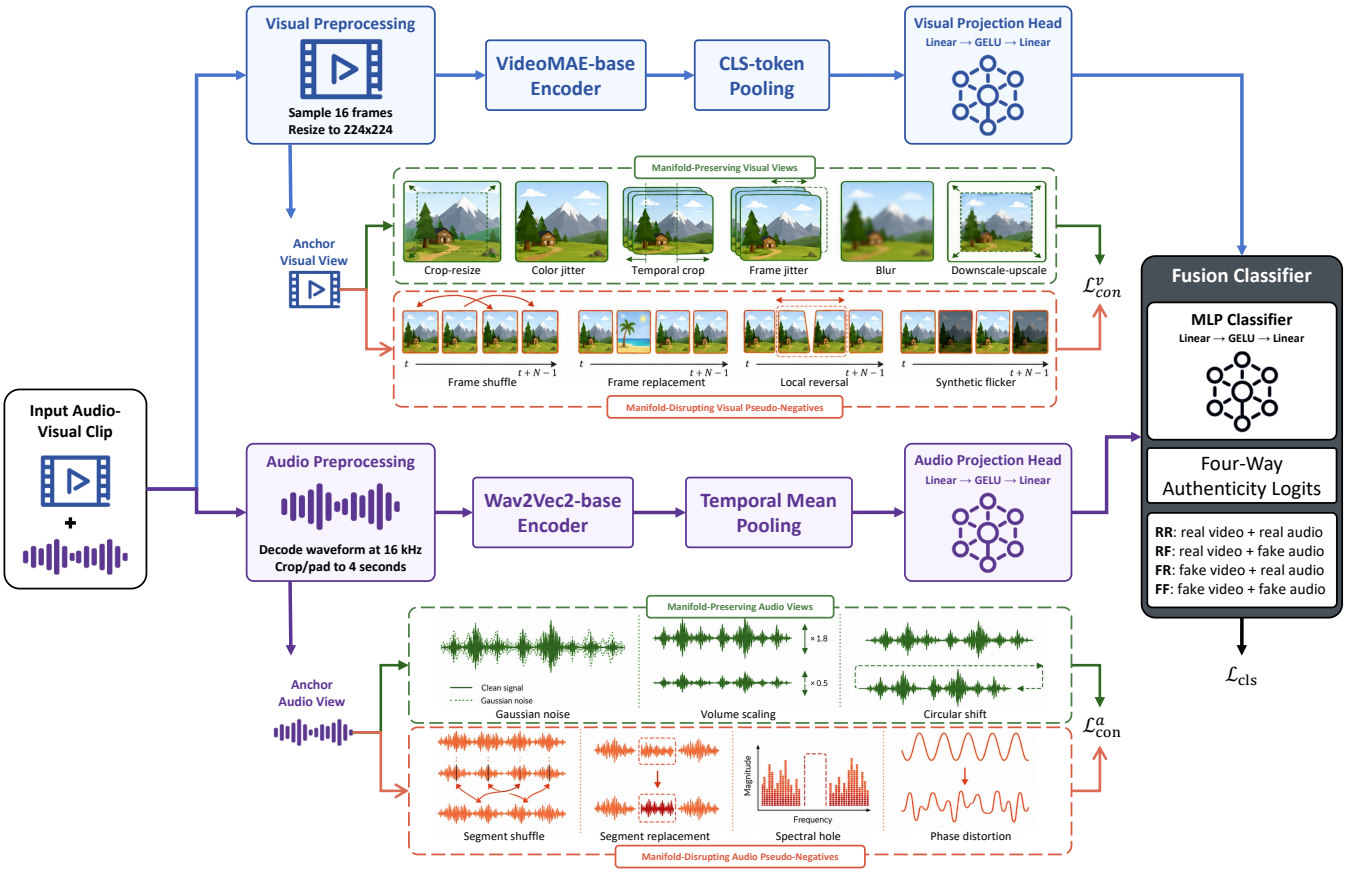

Figure 1: Overview of MVADETECTOR. Given an input audio-visual clip, the visual and audio streams are processed by VideoMAE-base and Wav2Vec2-base encoders, respectively. Each modality is projected into a 256-dimensional embedding space and optimized with modality-specific manifold contrastive learning. Manifold-preserving augmentations produce positive views, while manifold-disrupting augmentations produce structure-disrupted pseudo-negative views for video-side and audio-side contrastive regularization. The two modality embeddings are finally combined by a lightweight late-fusion classifier to predict four-way authenticity logits.

contrastive learning to regularize the visual and audio representations independently.

For each modality $r \in \{v, a\}$, we construct a triplet consisting of an anchor view, an authenticity-preserving positive view, and a structure-disrupted *pseudo-negative* view. The anchor is obtained from the original input. The positive view is generated by a *manifold-preserving* augmentation that keeps the modality on the same authenticity manifold, such as mild visual transformations or weak audio perturbations. In contrast, the pseudo-negative view is generated by a *manifold-disrupting* augmentation that intentionally disrupts temporal or spectral coherence within the same modality. Notably, this pseudo-negative view is not treated as a sample from another semantic class; instead, it serves as an *off-manifold* reference that encourages the encoder to capture authenticity-relevant structural regularities.

The contrastive objective is implemented as a cosine triplet loss [Schroff *et al.*, 2015]:

$$\mathcal{L}_{\text{con}}^{r} = \max\left(0, m - \cos(z_r, z_r^{+}) + \cos(z_r, z_r^{-})\right), \quad (4)$$

where z_r, z_r^{+}, and z_r^{-} denote the embeddings of the anchor, positive, and pseudo-negative views of modality r, respectively. The margin m controls the desired separation between positive and pseudo-negative views. All embeddings are ℓ_2-normalized before computing cosine similarity. We apply this objective separately to the visual and audio branches, yielding $\mathcal{L}_{\text{con}}^{v}$ and $\mathcal{L}_{\text{con}}^{a}$.

For video, temporal ordering and frame-level continuity are important authenticity-related cues. Therefore, temporal crop and frame jitter are regarded as mild perturbations, while frame shuffle and local reversal directly disrupt temporal coherence. For audio, spectral continuity and segment ordering motivate the design of audio-specific transformations.

Table 1 summarizes the augmentation design. For the visual branch, positive augmentations preserve global content and temporal ordering, while pseudo-negative augmentations perturb frame-level temporal coherence. For the audio branch, positive augmentations retain speech content and coarse prosodic structure, while pseudo-negative augmentations disrupt segment ordering, spectral continuity, or phase consistency. This design encourages each encoder to learn representations that are invariant to nuisance variations but discriminative to modality-specific authenticity cues.

Modality	Preserving positives	Disrupting pseudo-negatives
Video	Crop-resize; color jitter; temporal crop; frame jitter; blur; downscale-upscale	Frame shuffle; frame replacement; local reversal; synthetic flicker
Audio	Gaussian noise; volume scaling; circular shift	Segment shuffle; segment replacement; spectral hole; phase distortion

Table 1: Augmentation design for modality-specific manifold contrastive learning. Positive views preserve authenticity-relevant structure, while pseudo-negative views disrupt temporal or spectral coherence.

2.6 Learning Objective

MVADETECTOR is trained with a composite objective that combines four-way authenticity classification, modality-specific contrastive regularization, and auxiliary binary supervision:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_v \mathcal{L}^v_{\text{con}} + \lambda_a \mathcal{L}^a_{\text{con}} + \lambda_b \mathcal{L}_{\text{bin}}. \tag{5}$$

Here, $\mathcal{L}_{\text{cls}}$ denotes the four-class classification loss, $\mathcal{L}^v_{\text{con}}$ and $\mathcal{L}^a_{\text{con}}$ are the visual and audio contrastive losses, and $\mathcal{L}_{\text{bin}}$ is a binary auxiliary loss. The coefficients λ_v, λ_a, and λ_b control the relative contributions of the three auxiliary terms.

The classification loss is implemented as class-balanced cross-entropy, following the general practice of loss reweighting for imbalanced recognition [Cui *et al.*, 2019]. To mitigate the highly imbalanced label distribution, we assign each class a square-root inverse-frequency weight, i.e., the weight of class c is proportional to $n_c^{-1/2}$ and normalized to have mean one. Compared with full inverse-frequency weighting, this strategy increases the contribution of rare classes, especially **FR**, while avoiding overly large gradients from underrepresented categories.

The binary auxiliary loss provides an additional authentic-versus-synthetic training signal. We treat **RR** as the authentic class and aggregate **RF**, **FR**, and **FF** as synthetic classes. Instead of introducing a separate binary classification head, we derive the binary synthetic logit directly from the four-class logits:

$$\ell_{\text{syn}} = \text{logsumexp}_{c\in\{\text{RF,FR,FF}\}}(\ell_c) - \ell_{\text{RR}}. \tag{6}$$

This formulation encourages coarse real-versus-synthetic separation while keeping the primary supervision focused on fine-grained modality-wise authenticity prediction.

3 Experiments

3.1 Dataset and Evaluation Protocol

The challenge dataset is based on DDL-GAV/MVAD [IJCAI-ECAI 2026 AI Safety Workshop Organizers, 2026; Hu *et al.*, 2026] and consists of audio-visual samples from four modality-wise authenticity combinations: **RR** (real video-real audio), **RF** (real video-fake audio), **FR** (fake video-real audio), and **FF** (fake video-fake audio). The official evaluation considers both binary authenticity detection and fine-grained four-way classification. For each task, the score is computed as a weighted combination of AUC, accuracy, average precision, and F1:

$$S = 0.2\text{AUC} + 0.3\text{ACC} + 0.3\text{AP} + 0.2\text{F1}. \tag{7}$$

The final score further combines the binary and four-way scores:

$$S_{\text{final}} = 0.4S_{\text{bin}} + 0.6S_{\text{four}}. \tag{8}$$

Since the challenge does not provide an official validation split, we construct a stratified hold-out subset from the training data for local model selection. Unless otherwise specified, all reported competition results are obtained from the official leaderboard evaluation.

3.2 Implementation Details

We implement MVADETECTOR with VideoMAE-base [Tong *et al.*, 2022] as the visual encoder and Wav2Vec2-base [Baevski *et al.*, 2020] as the audio encoder. Each video is represented by 16 uniformly sampled frames resized to 224×224, and each audio stream is converted to a 4-second waveform at 16 kHz. Both visual and audio projection heads output 256-dimensional embeddings. The late-fusion classifier is a two-layer MLP with a 256-dimensional hidden layer, GELU activation, and a dropout rate of 0.1. The proposed framework avoids computationally expensive cross-modal interaction modules and only introduces lightweight projection heads and a late-fusion classifier.

We optimize the model using AdamW [Loshchilov and Hutter, 2019] with a backbone learning rate of 1×10^{-5} and a head learning rate of 1×10^{-4}, where the heads include the projection heads and the fusion classifier. The weight decay is set to 1×10^{-4}, and gradients are clipped with a maximum norm of 1.0. The learning rate follows one epoch of linear warmup followed by cosine annealing [Loshchilov and Hutter, 2017]. The effective batch size is 96, obtained with a per-GPU batch size of 8, gradient accumulation over 3 steps, and 4 GPUs. Training is conducted with PyTorch DistributedDataParallel [Li *et al.*, 2020] and FP16 automatic mixed precision [Micikevicius *et al.*, 2018] on 4 NVIDIA A100 80GB GPUs.

For *modality-specific contrastive learning*, we set the triplet margin to $m = 0.2$. The loss weights are set to $\lambda_v = 0.35$, $\lambda_a = 0.35$, and $\lambda_b = 0.02$ for the visual contrastive loss, audio contrastive loss, and binary auxiliary loss, respectively. We select the checkpoint from epoch 1 for final submission according to local validation performance.

At inference time, the four-way logits are converted into class probabilities using softmax. The binary synthetic probability is computed as $1 - P(\textbf{RR})$, corresponding to the probability that at least one modality is synthetic. For distributed inference, we remove duplicated predictions introduced by distributed sampler padding before generating the final submission.

Task	AUC	ACC	AP	F1	**Score**
Binary	0.9372	0.8969	0.9609	0.8967	**0.9241**
Four-way	0.8769	0.8384	0.7105	0.6367	**0.7674**
Final	–	–	–	–	**0.8301**

Table 2: Official leaderboard performance of our best submission. The binary and four-way scores are computed from AUC, accuracy, average precision, and F1. The final score follows the official weighting of 40% binary score and 60% four-way score.

3.3 Leaderboard Results

Table 2 reports the official leaderboard performance of our best submission. MVADETECTOR achieves strong binary authenticity detection performance, with an AUC of 0.9372, an F1 score of 0.8967, and a binary score of 0.9241. The four-way setting is more challenging, as it requires identifying the manipulated modality rather than only detecting the presence of synthetic content. In particular, the underrepresented **FR** class makes modality-wise classification sensitive to class imbalance. Nevertheless, our *modality-specific* contrastive regularization and class-balanced training strategy lead to a four-way score of 0.7674 and a final competition score of 0.8301.

3.4 Ablation Analysis

We conduct development ablations to assess the contribution of key components in MVADETECTOR. Since the official test labels are unavailable, these observations are based on local validation and development experiments, and are used to guide model design rather than to form a fully controlled benchmark.

Contrastive regularization. Removing modality-specific triplet losses causes a consistent degradation of approximately 2–3 points in four-way score on our local validation set. This suggests that manifold contrastive learning provides useful regularization beyond standard class supervision. By contrasting authenticity-preserving views with structure-disrupted pseudo-negatives, the encoders learn temporal, spectral, and structural cues for modality-wise authenticity prediction.

Class-imbalance handling. Class-balanced training is critical for the four-way setting. Without class-balanced weights and sampling, the model tends to bias toward frequent classes and often misclassifies rare **FR** samples as **FF**. This failure mode is especially harmful under macro-averaged evaluation, where underrepresented classes contribute equally to the final four-way metrics. The square-root inverse-frequency weighting provides a practical compromise: it increases the contribution of rare classes while avoiding the instability of full inverse-frequency weighting.

Binary auxiliary supervision. The binary auxiliary loss provides a useful but delicate training signal. Removing $\mathcal{L}_{\text{bin}}$ decreases the binary score by about 0.5 points, suggesting that coarse authentic-versus-synthetic supervision helps separate fully authentic samples from those containing at least one manipulated modality. However, increasing its weight to 0.1 or 0.2 degrades four-way performance. This suggests that an overly strong binary objective biases the model toward coarse real-versus-synthetic discrimination and weakens fine-grained modality-wise classification. Therefore, we keep this auxiliary objective weak and use it only as a regularizer.

Pooling strategy and checkpoint selection. We also examine several implementation choices that affect the final leaderboard score. For the VideoMAE [Tong *et al.*, 2022] branch, CLS pooling slightly outperforms mean pooling over all tokens, improving the final score by approximately 0.3 points. In addition, later checkpoints do not necessarily yield better generalization. Although training until epoch 6 increases confidence on the training distribution, the final score decreases from 0.8301 to 0.812, suggesting overfitting to frequent classes and reduced robustness on rare categories. We therefore use the epoch-1 checkpoint for the final submission.

4 Discussion and Limitations

The main strength of MVADETECTOR lies in its simplicity and alignment with the task structure. By factorizing audio-visual authenticity prediction into modality-wise evidence modeling, the framework avoids unnecessary cross-modal interaction modules while remaining efficient and reproducible. This design is particularly suitable for the four-way label space, where each class corresponds to a specific combination of visual and audio authenticity.

Despite its effectiveness, MVADETECTOR has several limitations. First, it does not explicitly model audio-visual synchronization, lip motion, or cross-modal semantic consistency. These cues may be important for samples in which the visual and audio streams appear realistic in isolation but are inconsistent as a pair. Second, the current augmentation policy relies on task-specific prior knowledge. Although the positive and pseudo-negative transformations are motivated by temporal and spectral coherence, learned augmentation policies or hard negative mining strategies may provide stronger contrastive supervision. Future work may explore adaptive augmentation discovery. Third, local validation remains imperfect because the challenge does not provide an official validation split. As a result, validation performance may not fully reflect leaderboard generalization, especially for underrepresented categories such as **FR**.

5 Conclusion

We presented MVADETECTOR, a modality-specific manifold contrastive learning framework for general audio-visual AIGC detection. The method combines open-source VideoMAE [Tong *et al.*, 2022] and Wav2Vec2 [Baevski *et al.*, 2020] backbones, lightweight late fusion, modality-specific contrastive regularization, class-balanced training, and weak binary auxiliary supervision. This combination is well matched to the factorized four-way structure of the challenge and achieves a final leaderboard score of 0.8301. Future work will explore explicit cross-modal consistency modeling, adaptive pseudo-negative mining, and more robust validation strategies for rare forgery categories.

References

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

framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

[Cai *et al.*, 2022] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content-driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10. IEEE, 2022.

[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.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[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. *arXiv preprint arXiv:2512.00336*, 2026.

[IJCAI-ECAI 2026 AI Safety Workshop Organizers, 2026] IJCAI-ECAI 2026 AI Safety Workshop Organizers. Track 2: General aigc audio-video detection. https://ai-safety-workshop-ijcai2026.github.io/Track2.html, 2026. Accessed: 2026-07-03.

[Katamneni and Rattani, 2023] Vinaya Sree Katamneni and Ajita Rattani. Mis-avoidd: Modality invariant and specific representation for audio-visual deepfake detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1371–1378. IEEE, 2023.

[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.

[Li *et al.*, 2020] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. Pytorch distributed: Experiences on accelerating data parallel training. *Proceedings of the VLDB Endowment*, 13(12):3005–3018, 2020.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[Micikevicius *et al.*, 2018] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

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

[Muppalla *et al.*, 2023] Sneha Muppalla, Shan Jia, and Siwei Lyu. Integrating audio-visual features for multimodal deepfake detection. In *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–5. IEEE, 2023.

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

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

[Yu *et al.*, 2023] Cai Yu, Peng Chen, Jiahe Tian, Jin Liu, Jiao Dai, Xi Wang, Yesheng Chai, Shan Jia, Siwei Lyu, and Jizhong Han. A unified framework for modality-agnostic deepfakes detection. *arXiv preprint arXiv:2307.14491*, 2023.