

cc_rmws at ImageCLEF 2026 Deepfake Task: Multi-Feature Audio Fusion and High-Frequency Image Modeling

Cheng Chen¹, Xiaobing Zhou*,¹

¹*School of Information and Engineering, Yunnan University, Kunming 650500, China*

Abstract

This paper presents a dual-modal deepfake detection system for Subtask 2 of the ImageCLEF 2026 Deepfake Detection Task. The task requires binary classification of audio and image samples under unseen generation methods.

For the audio modality, we extract self-supervised WavLM-large features, cepstral LFCC features, and phase-based MGD features, and apply a multi-view fusion strategy with a dual-branch backend. For the image modality, we adopt an SBI-based online pseudo-forgery generation strategy and a Swin-Small-based detector enhanced with high-frequency residual modeling.

To improve robustness, we further introduce a boundary-guided high-frequency variant to emphasize facial blending regions.

Experimental results on the official hidden test set show that our system achieves scores of 0.5551 for audio and 0.5858 for image. The results indicate that image-based modeling benefits more from online pseudo-forgery generation, while audio detection remains challenging under unseen spoofing conditions.

Keywords

Deepfake detection, ImageCLEF Deepfake 2026, Dual-modal detection, SBI-based pseudo-forgery generation, High-frequency residual information

1. Introduction

In recent years, with advances in generative artificial intelligence, the realism of synthetic media, including audio and images, has continued to improve. Voice conversion, neural vocoders, diffusion models, and face generation and editing techniques have made deepfake content increasingly difficult to distinguish by human inspection alone. If such content is used to impersonate identities, manipulate public opinion, or spread misinformation, it may pose serious risks to financial security, social trust, and public governance. Therefore, detecting deepfake content generated by unseen methods has become an important research problem in multimedia forensics.

Subtask 2 of the ImageCLEF Deepfake 2026 Detection Task is designed around this problem [1, 2]. It requires systems to perform binary authenticity classification for both audio and image samples, determining whether each input is real or a deepfake. Unlike methods evaluated only on fixed datasets, the official test set of this task contains not only deepfake samples generated by the organizers, but also samples submitted by other teams in the generation task. Therefore, this task places greater emphasis on model generalization under unseen generation algorithms and unseen data sources.

Although deepfake detection methods have made considerable progress in recent years, limited generalization remains a major challenge [3, 4, 5, 6]. Many detectors perform well when the training and test data follow similar distributions, but their performance often drops when test samples come from new generators, unseen post-processing pipelines, or different compression conditions. This is mainly because detection models may learn forgery traces specific to certain generation methods, rather than stable cues that generalize across different generators. In image-based detection, this may appear as dependence on specific texture patterns; in audio-based detection, it may appear as dependence on artifacts from particular speech synthesis systems or vocoders.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

✉ 12025215210@stu.ynu.edu.cn (C. Chen); zhoubx@ynu.edu.cn (X. Zhou*)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To address this issue, we built a dual-modal detection system for the ImageCLEF Deepfake 2026 Detection Task. The system processes audio and image inputs separately, with each modality trained and submitted independently. Instead of optimizing for a specific known forgery algorithm, this work focuses on the generalization problem under unseen generation methods. We therefore construct a participating system with separate audio and image branches, and analyze its performance across different subsets of the official evaluation data.

For the audio modality, we extract three types of features in parallel, namely WavLM-large self-supervised speech representations [7], Linear Frequency Cepstral Coefficients (LFCC) [8], and Modified Group Delay (MGD) features [9], and fuse them using a Triple-Fusion strategy. The fused features are then fed into a lightweight Dual-Branch back-end to jointly model long-range speech consistency and local acoustic anomalies. For the image modality, we start from real face frames in FaceForensics++ (FF++) c23 [10] and adopt a Self-Blended Images (SBI)-based online pseudo-forgery generation strategy [11] to construct training samples. This strategy does not rely on pre-generated deepfake videos. Instead, it dynamically generates pseudo-forged samples during training based on real faces and landmarks, together with color, geometric, texture, and frequency perturbations, so that the training samples vary in location, shape, and appearance. For detection, we utilize an HF-Swin structure based on Swin-Small [12] to jointly model RGB semantic features and high-frequency residual information. In addition, we implement a Boundary-guided HF variant, which applies boundary-aware spatial reweighting before feeding the high-frequency residual into the HF branch, encouraging the model to focus more on abnormal responses around facial boundaries and blending transition regions.

The main work of this paper can be summarized as follows:

- **Multi-feature fusion framework for the audio modality.** We build an audio detection system that combines WavLM-large, LFCC, and MGD features. Through the Triple-Fusion strategy and the Dual-Branch back-end, the system jointly models forgery cues from semantic, spectral, and phase-related perspectives. This design reduces the reliance on a single speech representation and allows us to observe its behavior under unseen audio forgery scenarios.
- **SBI-based image training strategy and HF-Swin detection framework.** For the image modality, we adopt an SBI-based online pseudo-forgery generation strategy to dynamically construct forged training samples from real face frames, reducing reliance on a fixed manipulated dataset. On the detection side, HF-Swin combines RGB backbone features with high-frequency residual features, allowing the model to use both global facial structure information and local texture anomalies. We also implement a Boundary-guided HF variant to strengthen the attention of the HF branch to facial boundaries and blending transition regions.
- **Dual-modal empirical analysis on the official hidden test set.** We submit both audio and image detection results to the official hidden test set of ImageCLEF Deepfake 2026, and analyze the system’s behavior across different subsets, including real samples, organizer-generated deepfakes, and participant-generated deepfakes. The official evaluation results show that our system achieves final scores of 0.5551 and 0.5858 on the audio and image tasks, respectively. Further analysis shows that the audio model performs well on real speech but remains sensitive to participant-generated unseen spoofed speech, while the image model shows relatively more balanced performance.

2. Related Work

2.1. Audio Representation Learning for Anti-Spoofing

Early audio anti-spoofing methods mainly relied on handcrafted acoustic features, such as LFCC and CQCC [8, 13]. These features are commonly used to describe spectral structures and local acoustic variations in speech, and they provide relatively strong interpretability in traditional speech spoofing detection tasks. Since many speech synthesis or conversion methods leave artifacts in spectral details, low-level acoustic features remain an important baseline for audio deepfake detection.

In recent years, self-supervised learning (SSL) speech models have been widely used for audio spoofing detection. Models such as WavLM, HuBERT, and wav2vec 2.0 learn context-dependent speech representations from large-scale speech data and capture long-range temporal dependencies [7, 14, 15]. These representations perform well in many standard evaluation settings and are useful for modeling speech content, prosodic structure, and global speech consistency.

However, relying only on self-supervised speech representations has limitations, especially when spoofing detectors are evaluated on out-of-domain or real-world audio data [6]. These models often encode rich semantic, speaker-related, and channel-related information, which may introduce dataset bias into the detector. When test samples come from new speech synthesis systems, unseen vocoders, or different compression conditions, high-level semantic representations may fail to capture stable forgery traces. In addition, artifacts from modern speech generation methods may appear in spectral details, phase structures, or short-term local variations, which may not be fully preserved by a single SSL representation.

Recent studies have therefore explored multi-view audio representations and feature fusion for spoofing detection [16, 17]. Compared with using a single feature, combining SSL representations, spectral features, and phase-related features can describe audio samples from different perspectives. The audio branch in this work follows this idea by combining WavLM-large, LFCC, and MGD to jointly model global speech representations, spectral variations, and phase cues, thereby reducing dependence on a single feature space.

2.2. Data Generation and High-Frequency Modeling for Image Forgery Detection

In image deepfake detection, the way training data is constructed directly affects model generalization. If a detector is trained only on samples generated by specific manipulation methods, it may learn generator-specific local artifacts or frequency traces [3, 18, 19]. When test samples come from new forgery methods, these cues may no longer be reliable, leading to degraded detection performance. To alleviate this problem, some methods construct more diverse training samples through data synthesis.

Self-Blended Images (SBI) is a representative online pseudo-forgery generation strategy [11]. Instead of relying on pre-generated deepfake videos, it starts from real face images and dynamically constructs and blends forged regions during training. This encourages the model to focus more on transition inconsistencies between real and forged regions, rather than memorizing texture patterns from a fixed generator. The image branch in this work adopts an SBI-based online pseudo-forgery generation strategy and combines color changes, geometric transformations, texture perturbations, and frequency perturbations during training to increase the diversity of training samples.

In addition to data generation, frequency-domain information has also been widely explored for image deepfake detection, since forgery artifacts may appear as subtle texture or spectral inconsistencies [5, 18, 19]. Although many forged samples appear visually realistic at the global level, subtle anomalies may still exist in local textures, boundary transitions, or frequency distributions. Therefore, detectors can introduce a high-frequency branch to capture local forgery traces that may not be directly captured by the RGB backbone. Compared with relying only on global semantic features, high-frequency residual information focuses more on local texture variations and edge discontinuities.

However, high-frequency features can also be affected by noise, compression artifacts, and background textures. If the model uses high-frequency responses from the entire image without constraint, it may treat irrelevant noise as forgery evidence. Recent works also show that high-quality deepfakes often contain subtle and localized artifacts, and that explicitly modeling artifact-prone regions can improve generalization [3, 4]. To address this issue, some methods use spatial constraints or gating mechanisms to control the effective range of high-frequency features. The Boundary-guided HF variant in this work follows this direction. It does not introduce an additional boundary supervision loss, but applies spatial reweighting before the high-frequency residual is fed into the HF branch, reducing the influence of irrelevant background high-frequency responses.

Overall, generalization in image deepfake detection is closely related to both training data construction and feature modeling. On the one hand, online pseudo-forgery generation can reduce the model’s

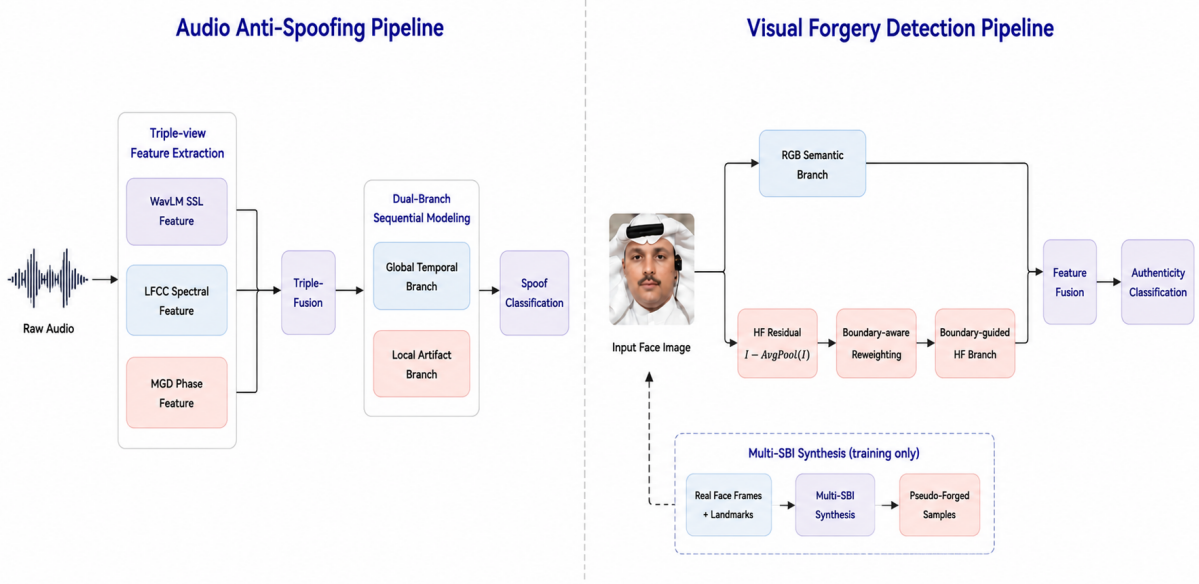


Figure 1: Overall framework of the proposed dual-modal deepfake detection system. The framework processes audio and image streams independently. The audio branch extracts WavLM, LFCC, and MGD features and performs Triple-Fusion followed by dual-branch sequential modeling. The image branch uses Multi-SBI for online pseudo-forgery generation during training, and combines RGB semantic features with boundary-guided high-frequency residual features for final classification.

reliance on fixed generators. On the other hand, combining RGB semantic features with high-frequency residual features allows the model to use both global structure and local forgery traces. The image branch in this work is built along these two directions: it uses an SBI-based online training strategy to increase sample diversity, and combines HF-Swin with Boundary-guided high-frequency modeling to enhance the perception of local anomalous regions.

3. Methodology

3.1. Overall Framework

The dual-modal framework constructed in this work is designed for audio and image deepfake detection under unseen manipulation techniques. The overall framework is shown in Figure 1.

For the audio modality, we adopt a multi-view acoustic feature extraction and temporal modeling strategy to capture forgery traces from semantic, spectral, and phase-related perspectives. Specifically, we extract WavLM self-supervised features, Linear Frequency Cepstral Coefficients (LFCC), and Modified Group Delay (MGD) phase features in parallel. These complementary feature representations are then fused and fed into a dual-branch temporal modeling module for the final authenticity prediction.

For the image modality, the framework adopts an SBI-based online generation strategy to dynamically construct forged images from real face samples, reducing the model’s reliance on fixed deepfake generators. For detection, we use an HF-Swin structure based on Swin-Small to jointly model RGB semantic features and high-frequency residual features. In addition, we implement a Boundary-guided HF variant, which applies boundary-aware spatial weights before the high-frequency residual enters the HF branch, enabling the HF branch to focus more on abnormal responses around facial boundaries and blending transition regions.

3.2. Audio Deepfake Detection

Traditional audio deepfake detection methods often rely on a single acoustic representation, which makes it difficult to model both high-level semantic information and fine-grained forgery artifacts. To

address this limitation, we adopt a multi-view acoustic feature extraction strategy that jointly models semantic, spectral, and phase-related information.

First, a pretrained WavLM front-end is used to extract self-supervised speech features [7], which capture high-level semantic abstractions and contextual dependencies. Second, LFCC features are extracted to describe the overall spectral structure [8]. In addition, MGD phase features are introduced to model subtle phase discontinuities [9] that may be produced by modern neural vocoders.

These three representations are first projected into a unified feature dimension and then fused by the Triple-Fusion module with frame-wise weighted fusion. The fused temporal representation is further fed into a dual-branch temporal modeling module. The global temporal branch learns long-range speech consistency, while the local artifact branch focuses on short-term acoustic anomalies and local forgery traces. Finally, the aggregated representation is passed to the classification head to produce the authenticity prediction.

3.3. Image Deepfake Detection

Existing visual deepfake detection methods can easily rely on forgery patterns left by specific generators. When test samples come from unseen manipulation methods, this reliance may reduce the generalization ability of the model. To alleviate this problem, we adopt an SBI-based online generation strategy in the image modality to generate forged images from real face samples [11]. Unlike directly using fixed deepfake videos, this strategy reads real face images and their landmark information during training, dynamically constructs local blending regions, and generates forged samples with color perturbations, geometric transformations, texture perturbations, and frequency perturbations. In this way, the forged regions in the training samples vary in location, shape, and appearance, reducing the model’s reliance on a single forgery pattern.

For the detection model, we use an HF-Swin structure to jointly model RGB semantic information and high-frequency residual information. The RGB branch uses Swin-Small as the backbone to extract global facial structure and semantic features [12], while the high-frequency branch supplements local texture anomalies, edge discontinuities, and fine-grained artifacts around blending regions. For an input face image I , we first obtain a smoothed image through average pooling and compute the difference between the original image and the smoothed image as the high-frequency residual:

$$F_{hf} = I - \text{AvgPool}(I). \quad (1)$$

Here, $\text{AvgPool}(\cdot)$ denotes the average-pooling smoothing operation, and F_{hf} denotes the high-frequency residual information in the input image. As shown in Eq. (1), this residual highlights the difference between the input image and its smoothed version. Compared with relying only on RGB semantic features, the high-frequency residual can more directly highlight local texture changes, edge anomalies, and frequency discontinuities [5, 19, 18], serving as a complement to the RGB backbone features.

In the Boundary-guided variant, we further introduce boundary-aware spatial reweighting before the high-frequency residual is fed into the HF branch. Specifically, a radial boundary weight W_b is constructed based on normalized image coordinates, encouraging the HF branch to focus more on the inner face region and boundary transition areas while reducing the interference from external background regions. The reweighted high-frequency input is formulated as:

$$\tilde{F}_{hf} = F_{hf} \odot W_b, \quad (2)$$

where $\odot$ denotes element-wise multiplication, and W_b denotes the boundary-aware spatial weight. As shown in Eq. (2), the boundary-aware weight reweights the high-frequency residual before it enters the HF branch. It should be noted that this boundary design is not an additional boundary classifier and does not introduce an independent boundary supervision loss. Instead, it is used as an input reweighting mechanism for the HF residual branch. In other words, it mainly changes where the high-frequency branch focuses spatially, rather than changing the overall classification objective.

Finally, the semantic features from the RGB backbone and the local features from the high-frequency branch are normalized separately and concatenated along the feature dimension. The concatenated features are then fed into the fusion classification head to produce the final real/fake prediction. In this way, the model can use both the global facial semantic information extracted by Swin-Small and the local forgery traces captured by the high-frequency residual branch, improving detection stability under unseen image forgery scenarios.

3.4. Implementation Details

The audio and image models are trained independently and submitted as two separate detection systems. All experiments are conducted on a single 32GB vGPU.

For the audio branch, the model takes three segment-level feature streams as input, including WavLM-large features, Linear Frequency Cepstral Coefficients (LFCC), and Modified Group Delay (MGD) features. The WavLM-large front-end is kept frozen during training to reduce overfitting on the proxy training data. The extracted features are projected into the same hidden dimension and fused by the Triple-Fusion module. The fused sequence is then processed by the Dual-Branch back-end, where one branch models global temporal consistency and the other branch focuses on local acoustic artifacts.

The audio model is trained with the Adam optimizer, using a learning rate of 1×10^{-4} , a batch size of 32, and 50 training epochs. The learning rate scheduler is ExponentialLR with a decay factor of 0.95. The classification head adopts Weighted Additive Angular Margin loss (WAAM) [20], which introduces an angular margin and class weighting to improve binary discrimination under class imbalance. In our setting, the WAAM margin is set to 0.4, the scale is set to 35, and the class weights are set to [0.1, 0.9]. The model is optimized together with the cross-entropy objective. During training, automatic mixed precision (AMP), gradient clipping, and feature-level augmentation are enabled. The augmentation includes time masking and channel masking inspired by SpecAugment [21].

For the image branch, the detector follows a two-branch architecture. The RGB branch uses Swin-Small pretrained on ImageNet-22K as the main backbone, while the high-frequency branch receives the residual input computed by $I - \text{AvgPool}(I)$. In the Boundary-guided variant, this residual is further reweighted by the spatial boundary weight before entering the high-frequency branch. The RGB branch outputs 768-dimensional features, and the high-frequency branch outputs 128-dimensional features. The two feature vectors are normalized separately, concatenated, and fed into the final binary classification head.

The image model is trained with the AdamW optimizer, using a learning rate of 1×10^{-4} , a batch size of 48, and 60 training epochs. The input image size is 224×224 . The learning rate scheduler is MultiStepLR, with milestones at epochs 10, 20, 35, and 50 and a decay factor of 0.5. The training objective consists of BCEWithLogitsLoss and a consistency regularization term, where the consistency loss weight is set to 0.25. During training, AMP is enabled, and the Swin-Small backbone is frozen for the first 3 epochs before being unfrozen for further training. Image augmentation includes SBI-based online generation, geometric transformations, color perturbations, and degraded-view perturbations.

4. Experiments

4.1. Audio Modality

4.1.1. Dataset and Evaluation Protocol

For the audio modality, we use ASVspoof 2019 LA as the main development dataset [22]. This dataset contains bona fide and spoofed speech under the logical access scenario, and provides three standard splits: train, development, and evaluation. In our experiments, the train split is used for model training, the development split is used for model selection and parameter tuning, and the evaluation split is used as an internal benchmark to check the basic behavior of the audio detector.

ASVspoof 2019 LA is used as a proxy training dataset in this work. Its distribution is not fully aligned with the official ImageCLEF 2026 Deepfake Detection test set, especially because the official test set

contains samples generated by unseen systems and by other participating teams. Therefore, the audio results should be interpreted mainly as a generalization test from a public anti-spoofing dataset to the official ImageCLEF evaluation data. During final submission, the audio model is not fine-tuned on the official test set. The selected checkpoint is directly used to infer binary labels for the official test audio files.

4.1.2. Training and Submission Settings

All input speech is first resampled to 16 kHz. During feature extraction, a fixed-duration segmentation strategy is used. Each utterance is divided into segments of 0.2 seconds, with a maximum utterance length of 6 seconds. Therefore, each utterance contains at most 30 temporal segments. For each segment, we extract three types of features: WavLM, LFCC, and MGD. WavLM provides self-supervised speech representations, LFCC describes spectral structure variations, and MGD supplements phase-related information.

To reduce the risk of overfitting caused by end-to-end fine-tuning on limited training data, the WavLM front-end is kept frozen during training. Only the back-end feature fusion module and the dual-branch detection network are trained. The final submitted audio model uses a WAAM classification head and is optimized together with the cross-entropy objective. In WAAM, the margin is set to 0.4, the scale is set to 35, and the class weights are set to $[0.1, 0.9]$, which helps alleviate the effect of class imbalance between real and spoofed speech samples.

To improve robustness to variations in speech duration and local acoustic perturbations, random temporal cropping, feature perturbation, and SpecAugment-like time and channel masking are applied during training. Model selection is mainly based on performance on the development split. The selected checkpoint is then used for inference on the official ImageCLEF audio test set, and the predictions are formatted according to the official submission requirements.

4.2. Image Modality

4.2.1. Dataset and Evaluation Protocol

For the image modality, we use real face videos from FF++ c23 as the training source [10]. Instead of directly training on a fixed set of manipulated videos, we construct online pseudo-forged samples from real face frames during training. This setting is intended to reduce the dependence on generator-specific artifacts and to expose the detector to more diverse blending patterns.

Following the official FF++ split, the internal development stage uses the train, validation, and test splits, which contain 720, 140, and 140 videos, respectively. During preprocessing, 256 frames are uniformly sampled from each training video, while 32 frames are sampled from each validation and test video for internal model selection and sanity checking. RetinaFace is used for face detection [23], and each detected face is cropped with approximately 1.3 times the bounding box to preserve facial boundary context. On the cropped face images, the dlib 81-point landmark model is used to extract facial landmarks [24]. Both the original landmarks and aligned landmarks are saved.

The preprocessing stage only stores real face images and their landmark information. No fixed forged images are generated in advance. During training, forged samples are generated online by the Multi-SBI strategy.

4.2.2. Training and Submission Settings

During image training, the model reads the preprocessed face images and their landmark information, and then generates corresponding forged samples online using Multi-SBI. Local blending regions are constructed based on facial landmarks, and forged images are generated with random boundary expansion, color perturbations, geometric transformations, texture perturbations, and high-frequency perturbations. Real samples and online-generated forged samples are combined into the same training batch and assigned binary labels.

The image detection model uses HFTopDownDetector with a Swin-Small backbone pretrained on ImageNet-22K. The input image size is 224×224 , the batch size is 48, and the model is trained for 60 epochs. The model contains an RGB main branch and a high-frequency residual branch. The RGB branch outputs 768-dimensional features, while the high-frequency branch obtains residual information through $I - \text{AvgPool}(I)$ and outputs 128-dimensional features. The two feature streams are concatenated after LayerNorm and then passed to the classification head to output a single logit.

The training objective uses BCEWithLogitsLoss with a consistency MSE loss weighted by 0.25. During official submission, the model outputs a sigmoid score for each test image and converts it into a hard label using the default threshold of 0.5, where real=0 and fake=1.

4.3. Official Evaluation Protocol

The official evaluation uses the hidden test set provided by the ImageCLEF 2026 Deepfake Detection Task organizers [2]. The task evaluates binary authenticity classification for both audio and image samples. Compared with a conventional closed test set, the official test data includes both organizer-generated samples and samples generated by other participating teams, which makes the evaluation focused on generalization under unseen generation methods and unseen data sources.

According to the official feedback, the results are divided into several subsets, including Baseline Deepfakes, Organizers Real Data, Organizers Real GT Data, Participant Deepfake Data, and Participant Deepfake Data Weighted. Baseline Deepfakes refers to basic forged samples generated by the organizers. Organizers Real Data and Organizers Real GT Data correspond to real samples provided by the organizers, with the latter being a further verified and hidden subset of real data. Participant Deepfake Data comes from forged samples submitted by other teams in the generation track.

Following the official evaluation protocol, each subset score is measured as binary classification accuracy. The final score is computed as a weighted sum of four subset accuracies:

$$S_{\text{final}} = 0.1S_{\text{base}} + 0.1S_{\text{real}} + 0.4S_{\text{realGT}} + 0.4S_{\text{partW}}. \quad (3)$$

Here, S_{base} , S_{real} , S_{realGT} , and S_{partW} denote the accuracies on Baseline Deepfakes, Organizers Real Data, Organizers Real GT Data, and Participant Deepfake Data Weighted, respectively. The unweighted Participant Deepfake Data score is reported for analysis, while the weighted participant score is used in the official final score.

4.4. Official Evaluation Results

This section reports the official evaluation results of our system on the hidden test set of ImageCLEF Deepfake 2026. Tables 1 and 2 show the subset accuracies and final weighted scores for the audio and image modalities, respectively.

Table 1

Official results of the audio detection system on the ImageCLEF Deepfake 2026 hidden test set.

Subset	Accuracy
Baseline Deepfakes	0.2979
Organizers Real Data	0.8146
Organizers Real GT Data	0.8625
Participant Deepfake Data	0.2407
Participant Deepfake Data Weighted	0.2472
Final Score	0.5551

Using the official weighting scheme in Eq. (3), the audio final score is computed as:

$$0.1 \times 0.2979 + 0.1 \times 0.8146 + 0.4 \times 0.8625 + 0.4 \times 0.2472 = 0.5551.$$

As shown in Table 1, the audio model performs relatively well on real speech data, achieving accuracies of 0.8146 and 0.8625 on Organizers Real Data and Organizers Real GT Data, respectively. This suggests that the model can identify the overall consistency patterns of real speech relatively stably. However, the accuracy drops to 0.2407 on Participant Deepfake Data and remains only 0.2472 on the weighted version. This indicates a clear distribution gap between the spoofed speech generated by other participating teams and the ASVspoof 2019 LA proxy dataset. It also suggests that the forgery cues learned from the public proxy dataset do not transfer well to all unseen generated speech. This reflects the continuing challenge of audio deepfake detection under unseen generators, vocoders, and different post-processing pipelines.

Table 2

Official results of the image detection system on the ImageCLEF Deepfake 2026 hidden test set.

Subset	Accuracy
Baseline Deepfakes	0.5266
Organizers Real Data	0.6491
Organizers Real GT Data	0.5948
Participant Deepfake Data	0.6363
Participant Deepfake Data Weighted	0.5758
Final Score	0.5858

For the image modality, the final score is computed using Eq. (3) as:

$$0.1 \times 0.5266 + 0.1 \times 0.6491 + 0.4 \times 0.5948 + 0.4 \times 0.5758 = 0.5858.$$

As shown in Table 2, the final score of the image modality is 0.5858, slightly higher than that of the audio modality. Across different subsets, the image model achieves an accuracy of 0.6363 on Participant Deepfake Data, which is higher than its performance on Baseline Deepfakes. This result suggests that the training strategy of constructing forged samples online from real images may help the model adapt to unseen image forgery methods. Since the forged images used during training do not come from a fixed generator, the model is less likely to memorize texture biases from a specific algorithm and is more likely to rely on blending boundaries, high-frequency anomalies, and local structural inconsistencies for prediction.

4.5. Ablation Study

To analyze the contribution of each component in our proposed framework, we conduct an ablation study on the key modules of both audio and image models. Since the system is designed for a shared evaluation benchmark with a fixed submission setting, we report controlled comparisons based on validation performance and model variants.

4.5.1. Audio Modality Ablation

We evaluate the contribution of different feature streams and fusion strategy. Removing WavLM features leads to a noticeable performance drop, indicating its strong contribution to robustness under unseen spoofing conditions. Similarly, removing LFCC or MGD features reduces performance, confirming that handcrafted spectral and phase information remains complementary to learned representations. Finally, replacing the Triple-Fusion module with simple concatenation leads to inferior performance, demonstrating the effectiveness of multi-view feature interaction.

4.5.2. Image Modality Ablation

We study the effect of key components in the image detector. Removing the high-frequency branch results in a noticeable performance drop, showing that residual-based frequency modeling is important

for detecting subtle forgery artifacts. Removing SBI-based online pseudo-forgery generation leads to a clear degradation in generalization ability, showing that diverse synthetic supervision is crucial for unseen manipulation robustness. Furthermore, disabling the boundary-guided weighting mechanism weakens the model’s ability to capture blending artifacts around facial edges.

Ablation summary. These results indicate that handcrafted acoustic features and self-supervised representations capture complementary information in the audio branch. In the image branch, frequency-aware modeling and online pseudo-forgery synthesis both contribute to generalization under unseen deepfake generation methods.

4.6. Analysis and Discussion

From the overall test results, the audio and image modalities show different generalization behaviors. The audio model achieves relatively high accuracy on real speech subsets, but its performance drops clearly on participant-generated spoofed speech. This suggests that although the Triple-Fusion structure can use SSL representations, spectral features, and phase-related features to describe audio samples from multiple perspectives, the gap between the training data and the test data still directly affects detection performance when the test samples come from completely unseen generation pipelines. In particular, current speech generation methods may produce more natural prosody, smoother spectral distributions, and more complex post-processing effects. Therefore, relying only on proxy training data such as ASVspoof 2019 LA is still insufficient to cover all possible spoofing patterns in the ImageCLEF hidden test set.

In contrast, the image model shows more balanced performance, especially with relatively higher accuracy on participant-generated forged images. This observation is generally consistent with the motivation of the SBI-based online training strategy. Since forged samples are dynamically generated during training, the model is exposed to forged regions with different locations, boundary shapes, and perturbation strengths. As a result, it is less likely to rely only on features from a fixed forgery method. Meanwhile, the high-frequency residual branch supplements local texture and frequency anomaly information. The Boundary-guided variant further applies boundary-aware spatial reweighting to the high-frequency residual, encouraging the model to focus more on abnormal responses around facial boundaries and blending transition regions, thereby improving its stability under unseen image forgery scenarios.

Nevertheless, the proposed system still has several limitations. First, the audio model is mainly trained on ASVspoof 2019 LA, whose distribution differs clearly from the participant-generated data in ImageCLEF 2026. Therefore, the audio branch still has limited performance under unseen generation domains. Second, although the image model improves its adaptability to unseen forged samples through the SBI-based online training strategy, the generated training samples are still mainly based on local self-blending of real faces. There may still be a gap between these samples and high-quality diffusion-based or end-to-end generated deepfakes. Finally, the audio and image models in this work are trained and submitted independently, and the complementary information between the two modalities has not yet been exploited. Future work may explore audio-image late fusion, test-domain adaptation, and data augmentation strategies that are closer to real generation processes to further improve system stability under unseen deepfake scenarios.

5. Conclusion

This paper presents our dual-modal detection system for the ImageCLEF 2026 Deepfake Detection Task. The system consists of two independent branches for audio and image detection, which are trained and submitted separately. For the audio modality, the system uses three types of features, namely WavLM-large, LFCC, and MGD, and performs binary classification through the Triple-Fusion strategy and the Dual-Branch back-end. For the image modality, the system starts from real face frames in FF++ c23, generates forged samples online with an SBI-based pseudo-forgery generation strategy, and uses

an HF-Swin model based on Swin-Small to jointly model RGB semantic features and high-frequency residual information. For the image branch, we also implement a Boundary-guided high-frequency residual weighting variant, which encourages the HF branch to focus more on abnormal responses around facial boundaries and blending transition regions.

According to the evaluation results on the official ImageCLEF test set, the audio and image branches show clear differences in performance. The audio model achieves relatively high accuracy on real speech subsets, but its performance drops clearly on participant-generated spoofed speech. This suggests that relying only on proxy data such as ASVspoof 2019 LA is still insufficient to cover the more complex unseen speech generation methods in the ImageCLEF hidden test set. In contrast, the image model shows more balanced overall performance, with more stable results on participant-generated forged images. This suggests that online pseudo-forgery generation based on real faces, together with the combination of RGB semantic features and high-frequency residual information, provides some adaptability to unseen image forgery methods.

Nevertheless, the evaluation results on the official ImageCLEF test set indicate that generalization to unseen deepfakes remains unresolved. The audio branch remains sensitive to unseen generators and unseen vocoders. Although the image branch is more stable than the audio branch, its forged training samples are still mainly based on self-blending, leaving a gap between these samples and deepfakes generated by high-quality diffusion-based or end-to-end generation models. In future work, the audio branch could incorporate data that better matches real generation scenarios, or explore test-domain calibration and lightweight SSL adapters. For the image branch, the effects of HF-Swin and Boundary-guided variants under different evaluation scenarios should be further investigated. In addition, since the audio and image models are currently submitted independently, future work could explore late fusion or confidence-based fusion to make use of complementary information between the two modalities.

Acknowledgments

We thank the ImageCLEF 2026 Deepfake Detection Task organizers for providing the evaluation platform, task resources, and official results.

Declaration on Generative AI

During the preparation of this work, the author used GPT for translation and writing assistance, and Codex for programming support. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

- [1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryilmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

- [2] D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, B. Ionescu, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, Overview of imageclef 2026 deepfake task: Multimodal detection and generation of deepfakes, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [3] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, D. Aouada, LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection, arXiv preprint arXiv:2401.13856 (2024).
- [4] D. Nguyen, M. Astrid, A. Kacem, E. Ghorbel, D. Aouada, Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection, arXiv preprint arXiv:2501.01184 (2025).
- [5] H. Kashiani, N. A. Talemi, F. Afghah, FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing, arXiv preprint arXiv:2509.22412 (2025).
- [6] A. Das, Y. El Kheir, C. Franzreb, T. Herzig, T. Polzehl, S. Möller, Generalizable audio spoofing detection using non-semantic representations, arXiv preprint arXiv:2509.00186 (2025).
- [7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, WavLM: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.
- [8] M. Sahidullah, T. Kinnunen, C. Hanilçi, A comparison of features for synthetic speech detection, in: Proceedings of Interspeech 2015, 2015, pp. 2087–2091.
- [9] H. A. Murthy, V. R. R. Gadde, The modified group delay function and its application to phoneme recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, 2003, pp. I–68.
- [10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, FaceForensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.
- [11] K. Shiohara, T. Yamasaki, Detecting deepfakes with self-blended images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18720–18729.
- [12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
- [13] M. Todisco, H. Delgado, N. Evans, A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients, in: Proceedings of Odyssey 2016: The Speaker and Language Recognition Workshop, 2016, pp. 283–290.
- [14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, HuBERT: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460.
- [15] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 12449–12460.
- [16] J.-W. Jung, H.-S. Heo, H. Tak, H.-J. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, N. Evans, AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, pp. 2405–2409.
- [17] J. Zhang, G. Tu, S. Liu, Z. Cai, Audio anti-spoofing based on audio feature fusion, Algorithms 16 (2023) 317. doi:10.3390/a16070317.
- [18] J. Li, H. Xie, J. Li, Z. Wang, Y. Zhang, Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6458–6467.
- [19] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, F3-Net: Frequency in face forgery network for video forgery detection, in: European Conference on Computer Vision, 2020, pp. 86–103.
- [20] J. H. L. Hansen, Z. Wang, Audio anti-spoofing using simple attention module and joint optimization based on additive angular margin loss and meta-learning, in: Proceedings of Interspeech 2022, 2022, pp. 376–380. doi:10.21437/Interspeech.2022-904.

- [21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, in: Proceedings of Interspeech 2019, 2019, pp. 2613–2617.
- [22] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech, *Computer Speech & Language* 64 (2020) 101114.
- [23] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, RetinaFace: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5203–5212.
- [24] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.