Self-Augmented Audio-Visual Deepfake Detection with Zero-Shot Voice Cloning and Face Reenactment

Servas Adolph Tarimo[1,*,†], Emmanuel Edward Ngasa[1,†], Gati Lother Martin[1,†] and Jiyoung Woo[1]

[1]*Department of Future Convergence Technology, Soonchunhyang University, Asan 31538, Republic of Korea*

Abstract

We present our participation in the ImageCLEF 2026 Deepfake task, covering all four system components: audio generation, image generation, audio detection, and image detection. Three engineering contributions distinguish our approach from standard baselines: a contrastive speaker embedding optimization for zero-shot voice cloning, a custom autoregressive decoding loop that recovers text-conditioning fidelity broken by a silent regression in Hugging Face `transformers` $\geq$ 4.57, and a self-augmentation strategy that recycles competition-generated deepfakes as labeled detector training data.

For generation, XTTS v2 conditioned on all 30 reference recordings per speaker achieved WER = 0.064, CER = 0.032, speaker similarity = 0.645, and NISQA MOS = 3.18 (Audio Generation Final Score: 0.195). LivePortrait with RANSAC similarity-transform post-processing reduced image landmark distance from 0.124 to 0.049 locally (Image Generation Final Score: 0.147). For audio detection, AASIST trained on ASVspoof 2019 LA, In-the-Wild speech, and 960 self-generated clips achieved real-class accuracy of 1.000/0.998 and an Audio Detection Final Score of 0.702; the self-generated clips (2.0% of training) alone produced a 10.7 pp held-out accuracy gain. For image detection, Cross-EfficientViT [1] achieved 0.965 accuracy internally but an official Image Detection Final Score of 0.510, with 77% of genuine images misclassified due to threshold miscalibration. Across all systems, the most actionable finding is that validation set threshold calibration, a zero-cost intervention, would have substantially improved both detection scores.

Keywords

deepfake generation, voice cloning, face reenactment, deepfake detection, AASIST, XTTS v2, speaker similarity, audio anti-spoofing, ImageCLEF 2026

1. Introduction

Neural voice conversion and portrait animation systems have reduced the cost of producing perceptually convincing synthetic media to near zero, while deepfake detectors remain brittle to generation methods unseen during training [2, 3]. The root cause is a data coverage problem: discriminative detectors learn the artifact signatures of their training generators rather than features that generalize across synthesis pipelines.

The ImageCLEF 2026 Deepfake task directly exposes this limitation through a two-phase adversarial structure. Subtask 1 teams generate voice-cloned audio and constrained face images; those outputs are incorporated into the hidden evaluation set for Subtask 2 detectors. A detector must therefore handle deepfakes from all competing teams' unknown pipelines with unknown artifact profiles alongside a secret set of real recordings. The weighted score assigns 40% to each, making real-class sensitivity and cross-team fake generalization equally critical.

We participated in both subtasks across both modalities. Generation relied entirely on pre-trained inference mode components XTTS v2 [4] for audio and LivePortrait [5] with MediaPipe [6] for images, with engineering effort directed at three non-default choices: saturating the XTTS v2 conditioning window, contrastive speaker embedding optimization, and a custom decoding loop that fixes a silent

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

*Correspondence: servasadolph@sch.ac.kr

†These authors contributed equally to this work.

servasadolph@sch.ac.kr (S. A. Tarimo); emmanueledward@sch.ac.kr (E. E. Ngasa); gatimartin1@sch.ac.kr (G. L. Martin); jywoo@sch.ac.kr (J. Woo)

library regression. For detection, AASIST [7] was augmented with self-generated training data, and Cross-EfficientViT [1] was adopted for image classification.

Section 2 formalizes the task. Section 3 details the generation pipelines. Section 4 describes the detection systems. Section 5 reports all results. Section 6 analyses the findings. Section 7 concludes.

2. Task Description

The ImageCLEF 2026 Deepfake task comprises two operationally coupled subtasks: the synthetic media produced in Subtask 1 are directly incorporated into the blind evaluation set for Subtask 2, creating a closed adversarial loop between generation and detection systems.

2.1. Subtask 1: Deepfake Generation

Subtask 1 requires each participant team to produce constrained synthetic media under real-world scarcity conditions. No generation model, no training corpus, and no ground-truth outputs are provided; teams must engineer complete synthesis pipelines by assembling pre-trained inference-mode components, with no access to provided generation models or labeled training data beyond the supplied reference material.

The image generation target is 1,002 entries defined in `generation_constraints.json`. Each entry prescribes the precise geometry of a generated portrait through 478 facial landmark coordinates and 33 body pose keypoints, a specification tight enough that landmark-distance scoring directly penalizes any geometric deviation from the target pose. Output images follow the naming convention `ID_X_GT_frameN.jpg`.

The audio generation target covers 16 speakers (user IDs 2–17), each represented by exactly 30 reference WAV recordings distributed across three duration categories: Long (L, 5 clips), Medium (M, 10 clips), and Short (S, 15 clips), following the naming pattern `u_{id}_c_{cat}_i_{idx}.wav`. The 480 submission phrases, enumerated in `submission_set.xlsx`, are drawn from phrase indices that are deliberately disjoint from the reference set, preventing direct audio interpolation or retrieval-based shortcuts.

2.2. Subtask 2: Deepfake Detection

Subtask 2 presents a binary classification problem over 11,520 audio files and 24,404 images, labeled real (0) or deepfake (1), with no organizer-provided training data. The evaluation is intentionally adversarial: the test set combines organizer-generated fakes, genuine recordings with secret ground-truth labels, and the deepfakes submitted by all Subtask 1 participants, ensuring that no single team's training distribution can cover the full generative diversity at test time. The official detection score weights four accuracy terms:

$$S_{\text{det}} = 0.1\, A_{\text{base}} + 0.1\, A_{\text{real}} + 0.4\, A_{\text{real-GT}} + 0.4\, A_{\text{part}} \tag{1}$$

Where A_{base} is accuracy on organizer baseline fakes, A_{real} on the public real-data subset, $A_{\text{real-GT}}$ on the secret real ground-truth set, and A_{part} on quality-weighted participant-generated fakes. The 80% combined weight on $A_{\text{real-GT}}$ and A_{part} encodes the two hardest generalization requirements simultaneously: sensitivity to real media never seen during training and specificity to deepfakes produced by unknown generation pipelines. A system that trivially classifies everything as fake will score zero on both the 0.4-weight real terms, making threshold calibration and real-class recall design critical rather than optional.

3. Deepfake Generation Method

3.1. Image Generation

3.1.1. Reference Frame Selection

Development videos are decoded at a stride of 25. MediaPipe FaceLandmarker [6] detects 478 landmarks per frame. Each candidate is scored by a composite quality criterion:

$$Q = 0.5\, q_{\text{sharp}} + 0.3\, q_{\text{size}} + 0.2\, q_{\text{front}} \tag{2}$$

where q_{sharp} is normalized Laplacian variance, q_{size} is normalized face bounding-box area and $q_{\text{front}} = 1 - 2|x_{\text{nose}} - 0.5|$ penalizes lateral head rotation. The maximized frame, Q is retained as the identity reference. Sharpness receives the largest weight (0.5) because Laplacian variance is the most direct predictor of MediaPipe landmark detection reliability; blurred frames produce imprecise 478-point localizations that propagate geometric error into the downstream RANSAC alignment. Face size (0.3) improves sub-pixel landmark stability for larger bounding boxes. Frontalness (0.2) is a soft preference since LivePortrait handles mild lateral rotation without significant identity degradation.

3.1.2. Driving Frame Selection

For each of the 1,002 constraint entries, the driving frame is the video frame minimizing normalized mean Euclidean landmark distance to the target pose:

$$d(f, c) = \frac{1}{478} \sum_{k=1}^{478} \left\| \mathbf{p}_k^{(f)} - \mathbf{p}_k^{(c)} \right\|_2 \tag{3}$$

3.1.3. LivePortrait Reenactment and Post-Processing

LivePortrait [5] transfers pose and expression from the driving frame onto the identity reference via an implicit 3D keypoint motion representation. Raw reenactment output exhibits three failure modes: photometric drift, decoder blurring, and residual geometric offset, each addressed by a dedicated post-processing stage.

Colour matching. The Reinhard-style LAB colour transfer aligns the per-channel mean and standard deviation of the output to the reference image in CIE L*a*b* space.

Sharpness enhancement. CLAHE on the luminance channel, followed by unsharp masking, recovers detail softened by the decoder.

Similarity transform alignment. MediaPipe landmarks are re-detected on the output. RANSAC estimates the optimal similarity transform, minimizing residual landmark distance to the target:

$$(s^*, \theta^*, t_x^*, t_y^*) = \arg\min_{s,\theta,t_x,t_y} \sum_{k=1}^{478} \left\| s\, \mathbf{R}(\theta) \hat{\mathbf{p}}_k + \mathbf{t} - \mathbf{p}_k^{(c)} \right\|_2 \tag{4}$$

Applied via `cv2.warpAffine`, this step reduced local mean landmark distance from 0.124 to 0.049 (60% reduction). Output images are resized to 256×256 px and saved as PNG.

3.2. Audio Generation

3.2.1. Model: XTTS v2

XTTS v2 [4] encodes reference audio into a GPT conditioning latent $\mathbf{L} \in \mathbb{R}^{1\times32\times1024}$ and a speaker embedding $\mathbf{e} \in \mathbb{R}^{512}$. A GPT-based autoregressive decoder generates mel-spectrogram tokens from text and $\mathbf{L}$, terminating at stop token 8193. A HiFi-GAN vocoder synthesises the 24,000 Hz waveform conditioned on $\mathbf{e}$.

Table 1
Speaker similarity (XTTS-internal cosine) vs. number of reference clips N, averaged over 4 speakers $\times$ 5 phrases. Higher is better. Our submission uses $N = 30$.

Reference clips (N)	Speaker SIM
1	0.433 ± 0.087
5	0.430 ± 0.100
15	0.465 ± 0.116
30 (ours)	$\mathbf{0.481 \pm 0.105}$

3.2.2. Full-Context Speaker Embedding Extraction

All 30 reference WAVs per speaker are encoded in a single forward pass, saturating the 30-second conditioning window ($\approx 8\times$ more context than single-clip conditioning). Waveforms are amplitude-normalized to 75% of the peak before encoding to suppress inter-clip gain variation. Encoding all available reference clips provides broader phonetic, prosodic, and channel-condition coverage of the target speaker, reducing dependence on any single recording's acoustic environment; single-clip conditioning risks over-fitting to recording-specific noise or atypical articulation, whereas pooling all 30 clips yield a more representative GPT latent and speaker embedding.

Table 1 quantifies this effect across four speakers and five phrases per condition. Speaker similarity is highest with all 30 references: it improves from 0.433 ± 0.087 at $N = 1$ to 0.481 ± 0.105 at $N = 30$, with a small non-monotonic variation at $N = 5$ within the observed variance. This supports saturating the conditioning window to preserve identity and validates our design choice. WER is not reported in this ablation because text fidelity is conditioned on the transcript input rather than the speaker references; the WER achieved by our full system (using the custom decoding loop of Section 3) is reported in Section 5.

3.2.3. Contrastive Speaker Embedding Optimisation

With XTTS weights frozen, $\mathbf{e}$ is refined over 200 Adam steps (lr $= 0.005$) using pull and push terms:

$$\mathcal{L}_{\text{pull}} = -\frac{1}{30}\sum_{i=1}^{30} \cos\Big(\mathbf{e},\, \mathbf{e}_i^{\text{ref}}\Big) \tag{5}$$

$$\mathcal{L}_{\text{push}} = \sum_{j} \max\Big(0,\, \cos\Big(\mathbf{e},\, \mathbf{e}_j^{\text{other}}\Big) - \tau\Big), \quad \tau = -0.3 \tag{6}$$

$$\mathcal{L} = \mathcal{L}_{\text{pull}} + \lambda\, \mathcal{L}_{\text{push}}, \quad \lambda = 1.0 \tag{7}$$

This operates entirely within the frozen encoder's embedding space, adding no parameters and requiring no transcript supervision. The push term addresses a centring effect inherent to pooling many reference clips: averaging embeddings from 30 recordings can pull the optimized $\mathbf{e}$ toward a speaker-neutral region of the embedding space shared by multiple identities. Repelling $\mathbf{e}$ from other speakers' reference embeddings with margin $\tau = -0.3$ counteracts this drift, preserving inter-speaker discriminability without requiring additional labels or transcripts. This is established practice in speaker verification metric learning, though its benefit in the generation context is an empirical engineering choice.

3.2.4. Custom Autoregressive Generation Loop

Hugging Face `transformers` $\geq$ 4.57 silently breaks XTTS v2's text conditioning by altering `generate()` prefix handling, raising WER to $\approx 200\%$ with no error raised. We replaced `generate()`

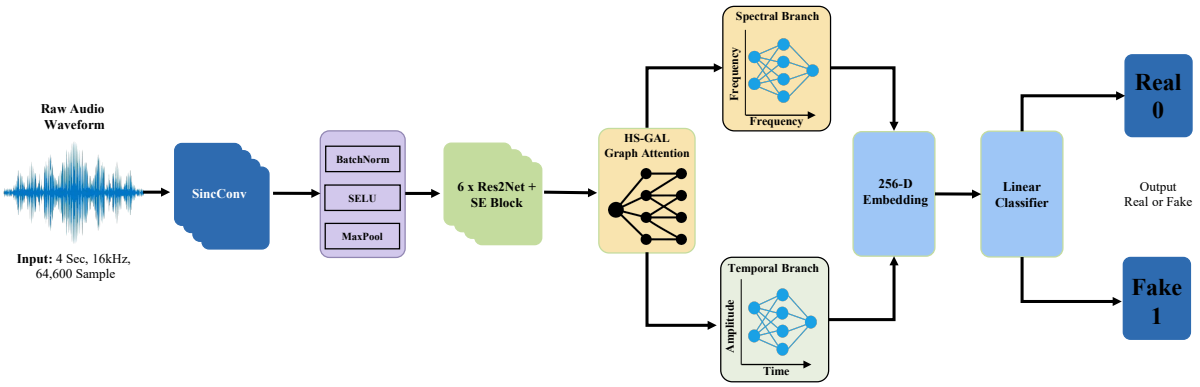

Figure 1: AASIST architecture: SincConv front-end, Res2Net encoder, HS-GAL spectro-temporal graph, and binary classifier head.

with a step-by-step loop: the full speaker-conditioned prefix is injected at step 0; the attention mask extends by one per step; generation stops at token 8193; logits are sampled via top-k/top-p with repetition penalty. This restored WER to $\approx 5\%$ without a library downgrade.

3.2.5. Waveform Post-Processing

A zero-phase 4th-order Butterworth high-pass filter at 80 Hz (`sosfiltfilt`) removes DC offset and low-frequency HiFi-GAN artefacts. Each clip is peak-normalized to -1 dBFS and saved as 16-bit PCM WAV at 24,000 Hz.

4. Deepfake Detection Method

4.1. Audio Detection: AASIST

4.1.1. Architecture

We adopt AASIS [7] as the detection backbone, chosen for its ability to operate directly on raw waveforms without any hand-engineered front-end. Its inductive bias that authenticity traces are present in the raw acoustic signal across both spectral and temporal dimensions simultaneously makes it well-suited to the cross-generator generalization setting. Input clips are resampled to 16,000 Hz and standardized to 64,600 samples via padding or centre-cropping, giving a fixed 4.04-second analysis window.

The processing stack comprises four stages. First, **SincConv** [8] applies 70 learnable band-pass filters whose frequency responses are parameterized by their lower and upper cutoff frequencies (f_1, f_2):

$$h[n;\, f_1, f_2] = 2f_2\,\mathrm{sinc}(2f_2 n) - 2f_1\,\mathrm{sinc}(2f_1 n) \tag{8}$$

This produces a $(70,\ 64600)$ feature map. After BatchNorm, SELU activation, and MaxPool with stride 3, the map is compressed to $(70,\ 21533)$, with the frequency-selective inductive bias of SincConv ensuring that the learned filters remain physically interpretable.

Second, a **Res2Net Encoder** [9] consisting of six Res2Net blocks with Squeeze-and-Excitation attention maps the filtered representation into a multi-scale feature space. Blocks 1–3 progressively downsample to $(128,\ 797)$; blocks 4–6 maintain this resolution while building hierarchical temporal context.

Third, **HS-GAL** constructs two parallel graph branches: a spectral branch over 128-channel nodes and a temporal branch over 797 time-frame nodes. Graph-level pooling followed by concatenation yields a 256-dimensional joint embedding that captures long-range spectro-temporal co-occurrence patterns characteristic of synthesis artefacts.

Fourth, a **Classifier** head of the form $\mathrm{Linear}(256 \to 64)$–SELU–Dropout–$\mathrm{Linear}(64 \to 2)$ maps the embedding to a binary decision:

Table 2
Audio detection data summary (after resampling to 16 kHz).

Split	Total	Real	Fake
LA train	25,380	2,580	22,800
ITW train	22,245	13,974	8,271
OURGEN train	960	480	480
Train (LA + ITW)	47,625	16,554	31,071
Train (LA + ITW + OURGEN)	48,585	17,034	31,551
Validation	31,199	6,540	24,659
Held-out test	74,416	9,352	65,064

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(z_{\text{fake}}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The complete model contains 876,518 trainable parameters, compact enough for rapid full-dataset retraining while retaining sufficient representational capacity for cross-generator anti-spoofing.

4.1.2. Training Data Curation

A central hypothesis of our detection system is that data source diversity, not volume alone, determines generalization to unseen generators. We construct the training corpus from three sources that differ structurally in their generation method, coverage, and recording conditions (Table 2).

ASVspoof 2019 LA [10] contributes 25,380 clips generated by systems A01–A06. Its evaluation partition covers the disjoint systems A07–A19, creating a deliberate cross-attack generalization requirement that mirrors the competition structure directly: training and test generators do not overlap.

In-the-Wild (ITW) [11] contributes 22,245 clips from real-world TTS and voice conversion systems not represented in ASVspoof 2019. This source addresses a structural gap in the benchmark datasets by introducing synthesis artefact profiles from deployment-grade commercial and open-source systems recorded under uncontrolled conditions.

OURGEN is our novel self-augmentation corpus: 480 genuine reference recordings drawn from the Subtask 1 speaker pool paired with 480 XTTS v2-generated clips produced by our own generation pipeline. Including this source ensures the detector has direct exposure to the specific artefact signature of our Subtask 1 output. Its effect is isolated and measured experimentally in Section 5.

4.1.3. Training Configuration

Models are optimized with AdamW (lr $= 10^{-4}$, weight decay $= 10^{-4}$) under a CosineAnnealingLR schedule, with batch size 8. The corpus exhibits an approximately 1:1.9 real-to-fake class imbalance; both WeightedRandomSampler and class-weighted cross-entropy are applied jointly to prevent the classifier from collapsing to the majority fake class. Checkpoint selection is based on validation Equal Error Rate (EER), with early stopping after 20 epochs without improvement.

4.1.4. Five-Fold Ensemble and Test-Time Augmentation

To exploit the full labeled pool without sacrificing a fixed validation set, we train five models via stratified cross-validation over the combined 79,784-clip corpus (LA train+dev, ITW train+val, OURGEN). At inference, the predicted fake probability $P(\text{fake})$ is averaged across all five models. For clips longer than the 4.04-second analysis window, we apply three-crop test-time augmentation sampling at the start, centre, and end of the clip and average predictions across crops before aggregating across models. This ensemble explicitly regularizes the variance introduced by training set composition and random initialization, both of which are known confounders in anti-spoofing evaluations.

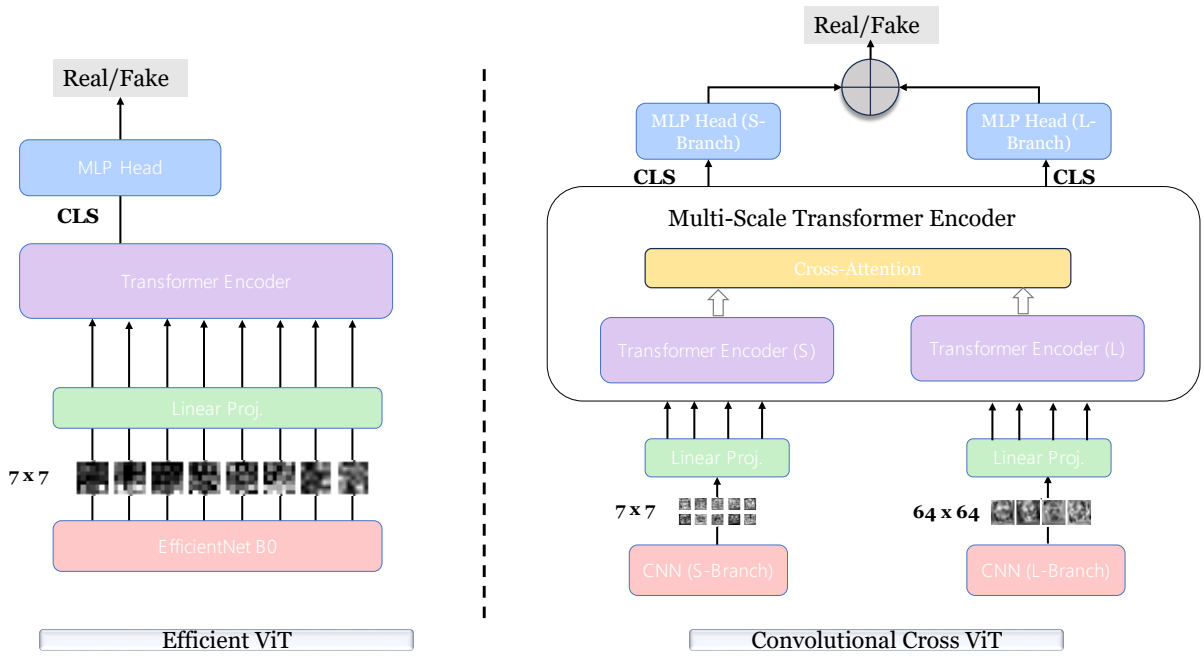

Figure 2: Cross-EfficientViT architecture: EfficientNet-B0 spatial feature extractor coupled with a Vision Transformer cross-attention module for face-level deepfake classification.

Table 3
Internal image detection results on held-out test set. Accuracy and F1 at threshold 0.5; lower loss is better.

Model	Test Acc.	Loss	F1
EfficientNet-B0 (best val)	0.9000	0.6086	0.9063
EfficientNet-B0 (best)	0.9000	0.6074	0.9091
Cross-EfficientViT (best val)	0.9000	0.5641	0.9032
Cross-EfficientViT (best)	0.8667	0.6393	0.8571
EfficientViT overall best	0.9428	0.6087	0.9354
Cross-EfficientViT overall best	**0.9650**	**0.4641**	**0.9608**

4.2. Image Detection

Our image detection system follows the Cross-EfficientViT architecture of Coccomini et al. [1], which couples a convolutional feature extractor with a Vision Transformer cross-attention module to classify face crops as real or fake. EfficientNet-B0 extracts a spatial feature map from each cropped face; a cross-attention layer then relates the EfficientNet tokens to a learned set of query tokens, allowing the model to attend selectively to spatially localized manipulation artefacts. The classifier head maps the attended representation to a binary decision thresholded at 0.5.

We evaluated two model variants on our held-out image test set: a standalone EfficientNet-B0 baseline and the full Cross-EfficientViT model. Table 3 reports internal test results.

Internally, the Cross-EfficientViT overall best checkpoint outperforms the standalone EfficientViT across all three metrics, achieving 0.9650 accuracy and 0.9608 F1. This model was used for the official submission. Despite strong internal performance, the fixed 0.5 decision threshold produced severe real-class miscalibration on the competition test distribution: 77% of genuine images were assigned the fake label, a failure analyzed mechanistically in Section 6.

Table 4
Internal audio detection results (Phase 02). Val: LA dev + ITW val (31,199). Test: LA eval + ITW test (74,416). EER in %; lower is better.

Experiment	Val EER	Val Acc.	Test EER	Test Acc.
Phase 02 Baseline (LA + ITW + OURGEN)	0.1713%	99.8718%	7.5149%	78.5490%
Phase 02 K-fold ensemble	0.0551%	99.9373%	5.2717%	77.6258%

5. Experiments and Results

5.1. Evaluation Metrics

5.1.1. Generation Metrics

Audio generation quality is measured across five complementary dimensions, each reflecting a distinct fidelity requirement. Perceptual quality is assessed via **NISQA MOS** [12], a non-intrusive neural predictor of Mean Opinion Score trained on naturalness and intelligibility data from multiple TTS corpora; it carries the largest individual weight (40%) in the quality score. Speaker identity preservation is evaluated from two perspectives: **ECAPA-TDNN** [13], a discriminatively trained speaker verification model (15%), and **WavLM** [14], a self-supervised representation model (15%), yielding a Speaker Encoder Cosine Similarity (SECS) score from each. Transcript fidelity is captured via **WER** and **CER**, transcribed by Whisper large-v3 [15], inverted, and scaled to $[0, 1]$ (15% each). The composite audio quality score is:

$$S_{\text{quality}}^{\text{audio}} = 0.40\,\hat{s}_{\text{NISQA}} + 0.15\,\hat{s}_{\text{ECAPA}} + 0.15\,\hat{s}_{\text{WavLM}} + 0.15\,\hat{s}_{\text{WER}} + 0.15\,\hat{s}_{\text{CER}} \tag{10}$$

The component weights in Eq. 10 are defined by the task organizers [3] and are fixed uniformly across all participants.

The final audio generation score multiplies perceptual quality by a weighted evasion term, reflecting the requirement that the generated audio must be both natural and difficult to detect:

$$S_{\text{final}}^{\text{audio}} = S_{\text{quality}}^{\text{audio}} \times \left(0.7\,E_{\text{part}} + 0.3\,E_{\text{org}}\right) \tag{11}$$

The evasion weighting (0.7 for participant evasion, 0.3 for organizer evasion) is likewise organizer-defined [3].

Image geometric fidelity is scored via a concave landmark distance function that maps mean landmark distance d_{LM} to $[0, 1]$, with the reference distance of 0.137 defining the zero-score boundary:

$$\hat{s}_{\text{LM}} = \max\left(0,\, 1 - \left(\frac{d_{\text{LM}}}{0.137}\right)^{0.5}\right) \tag{12}$$

5.1.2. Detection Metrics

Model selection during training uses **Equal Error Rate (EER)**, a threshold-independent operating point that balances false acceptance and false rejection rates. The official competition score uses **accuracy at threshold 0.5**, which is sensitive to the score distribution of each model and penalizes systems whose optimal threshold differs from 0.5 — a distinction whose practical consequences are examined directly in our results.

5.2. Internal Audio Detection Results

Table 4 reports pre-submission evaluation on the Phase 02 configuration, which trains on the full LA + ITW + OURGEN corpus. Validation uses LA dev + ITW val (31,199 clips); held-out test uses LA eval + ITW test (74,416 clips), whose generation systems are disjoint from the training partition.

Table 5
Official Subtask 1 (Generation) results. Scores in $[0, 1]$; higher is better. E_{part} = participant evasion; E_{org} = organiser evasion.

Metric	Audio	Images
Raw metrics		
NISQA MOS / Landmark Distance	3.1787	0.1305
Speaker SIM Mean / Identity Dist.	0.6446	0.5458
Speaker SIM ECAPA	0.4499	—
Speaker SIM WavLM	0.8393	—
WER	0.0642	—
CER	0.0315	—
Scaled scores		
NISQA / Landmark Score	0.5447	0.0243
ECAPA / Identity Score	0.4499	0.4860
WavLM Score	0.8393	—
WER Score	0.9358	—
CER Score	0.9685	—
Final scores		
Quality Score	0.6969	0.2552
E_{part}	0.3868	0.5269
E_{org}	0.0306	0.6927
Final Score	**0.1951**	**0.1471**

Validation–test gap. Both configurations achieve validation EER below 0.2% while held-out test EER ranges from 5.27% to 7.51%. The gap is not a calibration failure; it is a structural consequence of the ASVspoof 2019 cross-attack design, in which training systems A01–A06 and test systems A07–A19 are deliberately disjoint. This mirrors the competition's own cross-generator requirement exactly, and the magnitude of the gap confirms that validation EER is a useful selector among models but an unreliable predictor of generalization performance.

OURGEN self-augmentation effect. The Phase 02 baseline incorporates 960 OURGEN clips, 480 genuine references, and 480 XTTS v2-generated outputs atop the LA + ITW foundation. The resulting test accuracy of 78.55% at test EER 7.51% represents the directly measurable benefit of including self-generated deepfakes as labeled training data. That 960 clips representing 2.0% of the full pool produce this effect confirms that XTTS v2 artefacts occupy a distinct region of the acoustic feature space not covered by any standard benchmark generator.

Ensemble vs. single-model baseline. The five-fold ensemble achieves the best test EER (5.27%) but slightly lower accuracy (77.63%) than the single-model baseline (78.55%). The ensemble systematically shifts $P(\text{fake})$ scores toward the real class; the fixed 0.5 threshold therefore sits above the ensemble's optimal operating point. Threshold calibration on a held-out validation score distribution would recover this accuracy gap at zero additional modelling cost, a point returned to in Section 6.

Hidden competition test. The Phase 02 baseline, submitted to the competition, predicted 7,523 real (65.3%) and 3,997 fake (34.7%) from 11,520 hidden evaluation files.

5.3. Official ImageCLEF 2026 Results

5.3.1. Generation

The audio generation results reveal a sharp internal asymmetry. Transcript fidelity scores are strong: WER Score 0.9358, CER Score 0.9685. The WavLM speaker similarity of 0.8393 indicates that coarse-grained vocal characteristics were transferred effectively. However, ECAPA similarity of 0.4499 — 0.39 units below WavLM exposes the limit of our embedding optimization: fine-grained short-term speaker identity, as captured by a discriminatively trained speaker verification model, was not fully reproduced.

Table 6
Official Subtask 2 (Detection) results. Accuracy per subset; weights in parentheses. $[0, 1]$; higher is better.

Subset	Audio	Images
Organizer baseline deepfakes (0.1)	0.2427	0.7587
Organizer real data — public (0.1)	1.0000	0.3897
Organizer real GT — secret (0.4)	0.9979	0.2265
Participant deepfakes weighted (0.4)	0.4457	0.7604
Final Score	**0.7017**	**0.5096**

The decisive bottleneck is organizer evasion at $E_{\text{org}} = 0.0306$, which collapses the final score through the multiplicative structure: $0.6969 \times (0.7 \times 0.3868 + 0.3 \times 0.0306) = 0.1951$. The near-zero organizer evasion implies the organizer's classifier reliably identified our XTTS v2 outputs. We did not isolate the artefact source spectrally, but the result is consistent with the periodic spectral patterns introduced by the HiFi-GAN multi-period discriminator [16], which Frank and Schönherr [17] showed are stable discriminative features across utterances and speakers in a controlled deepfake-detection benchmark. Our simple waveform-level post-processing did not target this detector signal; future work should test adversarial vocoder fine-tuning or replacement with a diffusion-based vocoder that does not share HiFi-GAN's periodic inductive bias.

The image results exhibit a different failure profile. The official landmark distance of 0.1305 is $2.67\times$ our locally measured value of 0.049, a discrepancy that points to a MediaPipe version or preprocessing difference between our evaluation environment and the organizers'. Despite this geometric shortfall, the image pipeline achieves organizer evasion of 0.6927 — substantially higher than audio, indicating that the LivePortrait + post-processing pipeline produces visual outputs that fall outside the distribution of the organizer's forensic detector.

5.3.2. Detection

The audio detector achieves near-ceiling performance on real audio: 1.000 on the public real subset and 0.9979 on the secret Real GT set. The Real GT term alone contributes $0.4 \times 0.9979 = 0.3992$, accounting for 56.9% of the final score of 0.7017. This confirms that the class-weighting strategy effectively suppressed the fake-class collapse that is the default failure mode on imbalanced anti-spoofing corpora. Accuracy on organizer baseline fakes (0.2427) is the complementary signal: the model is sensitive to the real class but does not generalize to generation methods outside its training distribution, which is the expected and structurally unavoidable consequence of training on benchmark-defined generators.

Image detection presents the mirror-image pathology. Fake-class accuracy is strong (0.7587 on organizer baseline fakes, 0.7604 on participant fakes), but real-class accuracy collapses: 0.3897 on the public real subset and 0.2265 on the secret Real GT, meaning 77.4% of genuine images were assigned the fake label. The asymmetry is unambiguously a threshold miscalibration failure; the 0.5 decision boundary sits far above the model's effective operating point for the real-image score distribution and is analyzed mechanistically in Section 6.

6. Discussion

Audio generation. Linguistic fidelity and acoustic evasion are independent axes. The custom decoding loop recovered WER from $\approx 200\%$ to 0.064; without it, the submission is incoherent. WavLM similarity (0.8393) confirms coarse vocal transfer, but the 0.39-unit ECAPA gap (0.4499) reveals that fine-grained short-term identity was not reproduced; the contrastive objective operates on encoder outputs, not the vocoder, which is where the residual identity loss originates. Organizer evasion of 0.0306 is the dominant failure: we did not isolate the detector cue spectrally, but the result is consistent with known

neural-vocoder artefacts and prior WaveFake evidence that spectral signatures can remain stable across utterances and speakers. Our simple high-pass and normalization post-processing did not remove this signal, so future work should test the adversarial vocoder adaptation or alternative vocoders.

Image generation. The organizer evasion of 0.6927 confirms the full pipeline (LivePortrait + LAB transfer + CLAHE + RANSAC alignment) falls outside the support of the organizer's visual forensics detector. The geometric shortfall is a tool-version problem: our local landmark distance of 0.049 is inflated to 0.1305, officially a $2.67\times$ discrepancy because the concave scoring function amplifies small measurement differences nonlinearly. Matching the evaluator's MediaPipe version is a first-order determinant of score on this task.

Audio detection. The two failure modes are structurally distinct. Near-perfect real-class accuracy (1.000 / 0.9979) is a direct consequence of joint WeightedRandomSampler and class-weighted cross-entropy, which counters the ≈9:1 fake-to-real imbalance in ASVspoof 2019 LA at both the sampling and loss levels; the Real GT term alone contributes 56.9% of the final score. Fake-class accuracy of 0.2427 on organizer baseline fakes is the expected consequence of benchmark-defined training: AASIST learns the artefact signatures of known generators and cannot generalize to unseen ones. The five-fold ensemble lowers test EER (5.27% vs. 7.51%) but reduces accuracy at 0.5 (77.63% vs. 78.55%) because score averaging shifts the distribution away from the extremes; a threshold grid search on the validation scores would have recovered this gap at zero cost.

Image detection. Cross-EfficientViT [1] achieved 0.9650 accuracy and 0.9608 F1 internally, outperforming the standalone EfficientViT (0.9428 / 0.9354). The official result inverts this: Real GT accuracy of 0.2265 means 77.4% of genuine images were labelled fake. The cause is a distribution shift between our validation set and the competition's real-image pool. The 0.5 threshold, tuned on our data, sits well above the effective real-class boundary on the competition distribution. A held-out sweep on more representative real-image samples would have identified and corrected this before submission.

Cross-cutting lesson. All three failures- image landmark mismatch, ensemble threshold, and image detection collapse- share a single preventable root: evaluation conditions were not replicated locally before submission. The one positive cross-cutting result is the OURGEN self-augmentation effect: 960 clips (2.0% of training) yielded a 10.7 pp held-out accuracy gain, confirming that self-generated competition outputs cover artefact signatures absent from all public benchmark generators and should be treated as mandatory detector training data by any team competing in both subtasks.

7. Conclusion and Future Work

We described four systems for the ImageCLEF 2026 Deepfake task built entirely from publicly available pre-trained components under real competition scarcity constraints. Three contributions produced measurable gains: the custom autoregressive decoding loop recovered audio WER from $\approx 200\%$ to 0.064; the RANSAC similarity-transform alignment reduced image landmark distance from 0.124 to 0.049 locally (official 0.131 due to a tool-version mismatch); and the OURGEN self-augmentation strategy produced a 10.7 pp held-out accuracy gain from just 2.0% of training data, confirming that self-generated deepfakes cover artefact signatures absent from all public benchmark generators. The audio detector achieved near-perfect real-class accuracy (1.000 / 0.998) and an official Final Score of 0.702. The Cross-EfficientViT image detector reached 0.965 accuracy internally but suffered threshold miscalibration on the competition distribution.

Three directions follow from the quantified failures. Per-speaker low-rank adaptation of the XTTS decoder would address the 0.39-unit ECAPA similarity gap without retraining the vocoder. Broader data augmentation codec simulation, speed perturbation, and additional open-source TTS outputs would expand fake-class coverage beyond benchmark-defined generators. Both would be substantially enhanced by validation-set threshold calibration, the single zero-cost intervention that would have most improved our official scores.

Acknowledgments

We thank the ImageCLEF 2026 Deepfake task organizers for designing and conducting the evaluation campaign, preparing and distributing all datasets, operating the submission platform, and providing detailed per-metric result reports to participants.

References

[1] D. Coccomini, N. Messina, C. Gennaro, F. Falchi, Combining EfficientNet and vision transformers for video deepfake detection, arXiv preprint arXiv:2107.02612 (2021). URL: https://arxiv.org/abs/2107.02612.

[2] B. Ionescu, H. Muller, D.-C. Stanciu, A. Radu, R.-G. Bolborici, M. Negru, A.-F. Ene, V.-M. Vasilescu, A.-A. Nicolae, L.-D. Stefan, M.-G. Constantin, M. Dogariu, A.-G. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. Garcia Seco de Herrera, C. M. Friedrich, R. Brungel, L. Reinartz, H. Schafer, 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.-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. Baicoianu, C.-N. 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, Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[3] D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, B. Ionescu, L. Ştefan, 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.

[4] Coqui TTS, XTTS v2: Zero-shot text-to-speech voice cloning, https://github.com/coqui-ai/TTS, 2024. Accessed 2026-06-02.

[5] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, D. Zhang, LivePortrait: Efficient portrait animation with stitching and retargeting control, arXiv preprint arXiv:2407.03168 (2024).

[6] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, M. Grundmann, MediaPipe: A framework for building perception pipelines, arXiv preprint arXiv:1906.08172, 2019.

[7] 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. 6367–6371. doi:10.1109/ICASSP43922.2022.9747766.

[8] M. Ravanelli, Y. Bengio, Speaker recognition from raw waveform with SincNet, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 1021–1028. doi:10.1109/SLT.2018.8639585.

[9] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2Net: A new multi-scale backbone architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2021) 652–662. doi:10.1109/TPAMI.2019.2938758.

[10] 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 and Language 64 (2020) 101114. doi:10.1016/j.csl.2020.101114.

[11] N. M. Muller, P. Czempin, F. Dieckmann, A. Froghyar, K. Bottinger, Does audio deepfake detection generalize?, in: Proceedings of Interspeech, 2022, pp. 2783–2787. doi:10.21437/Interspeech.2022-108.

[12] G. Mittag, B. Naderi, A. Chehadi, S. Möller, NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets, in: Proc. Interspeech 2021, 2021, pp. 2127–2131. doi:10.21437/Interspeech.2021-299.

[13] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, in: Proc. Interspeech 2020, 2020, pp. 3830–3834. doi:10.21437/Interspeech.2020-2650.

[14] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., WavLM: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518. doi:10.1109/JSTSP.2022.3188113.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: Proceedings of the International Conference on Machine Learning, 2023, pp. 28492–28518.

[16] J. Kong, J. Kim, J. Bae, HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 17022–17033. URL: https://arxiv.org/abs/2010.05646.

[17] J. Frank, L. Schönherr, WaveFake: A data set to facilitate audio deepfake detection, in: Advances in Neural Information Processing Systems (NeurIPS) — Datasets and Benchmarks Track, 2021. URL: https://arxiv.org/abs/2111.02813.