Multi-Backbone Self-Supervised Ensembles for Audio Deepfake Detection and a Cross-Track Analysis of Generation–Detection Asymmetry

Notebook for the ImageCLEF at CLEF 2026

Seunghyun Kim[1,*], Junghyun Kim[1] and Jiyoung Woo[1]

[1]Soonchunhyang University, Department of AI and Big Data Engineering, Asan, Republic of Korea

Abstract

This paper describes the participation of team "Go-To-Germany" in the ImageCLEF 2026 Audio Deepfake Detection and Generation task. Our detection system, built on a four-backbone self-supervised learning (SSL) ensemble combining WavLM-Large, Wav2Vec2-XLS-R-300M, ECAPA-TDNN, and x-vector representations, achieved a final score of **0.9522** on the official ImageCLEF 2026 evaluation, with perfect accuracy (1.0000) on participant-generated deepfakes and 0.8875 on the held-out organizer ground-truth real data. For the Generation sub-task, our official team submission — an F5-TTS v1 baseline processed with a uniform reverberation pass, submitted as a deliberate anti-forensic probe — ranked first with a final score of 0.4304[1] (word error rate (WER) 4.99%, character error rate (CER) 2.07%); details of our four-model program (GLM-TTS, F5-TTS, XTTS v2, CosyVoice3), from which the official entry was drawn, appear in §3. We present a cross-track analysis revealing a pronounced asymmetry: our detection system identifies 100% of participant-generated deepfakes, while our official generation entry — despite ranking first in the Audio Generation sub-task and evading 61.4% and 56.2% of participant and organizer detectors — attains a Final Score of 0.4304 against 0.9522 on the Detection side. We further report falsification-based ablation experiments (LOSO 56-speaker cross-validation, three-region backbone geometry, bootstrap confidence intervals, and PCA analysis) that motivate our architectural-insurance hypothesis for multi-backbone SSL ensembling. We complement these results with five cross-track insights and five pre-registered falsification experiments connecting generation-side evasion to detection-side design decisions, and we openly report an 11.25% false-positive gap on held-out organizer real recordings as the principal open challenge for deployment.

Keywords

Audio Deepfake Detection, Speech Generation, Self-Supervised Learning, Multi-Backbone Ensemble, Text-to-Speech, ImageCLEF 2026

1. Introduction

1.1. Task Overview

The ImageCLEF 2026 Deepfake Task [1], part of the broader ImageCLEF 2026 evaluation campaign [2], comprises two complementary audio sub-tasks. The Detection sub-task asks whether a given audio file is a genuine recording or a synthesized deepfake; the evaluation set combines (i) deepfakes generated by the participating teams, (ii) baseline deepfakes provided by the organizers, and (iii) real recordings, with the official Final Score computed as a weighted average across four categories. The Generation sub-task asks each team to synthesize 480 utterances from sixteen target speakers; submissions are scored on audio quality (the Non-Intrusive Speech Quality Assessment (NISQA) [3], WER, CER, and speaker similarity) multiplied by their ability to evade both participant-built and organizer-built detectors.

The defining difficulty of both sub-tasks is *generalization*. On the detection side, the text-to-speech (TTS) systems used by competing teams are not disclosed in advance, so a detector that overfits to any particular generator family fails on unseen ones; this difficulty has been formalized in the ASVspoof series [4]. On the generation side, the detectors deployed by other participants and by the organizers

[1]By coincidence, the official Generation Final Score (0.4304) numerically equals the PCA-192 inter-region correlation reported in §6.3; the two are unrelated measurements.

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

*Corresponding author.

ksh011018@sch.ac.kr (S. Kim); jhyeun1234@sch.ac.kr (J. Kim); jywoo@sch.ac.kr (J. Woo)

are likewise undisclosed, so a generator that optimizes against any single detector family is unlikely to generalize.

1.2. Our Team's Position

We submitted to both sub-tasks under the same single-researcher compute budget (one A100 GPU), which makes our submissions an unusually clean controlled comparison of the same team's offensive versus defensive capability. During the Generation work we built six internal detectors (mel-frequency cepstral coefficients (MFCC) with deltas, linear-frequency cepstral coefficients (LFCC), mel-spectrogram statistics, spectral features, prosody, and raw-waveform statistics) to verify that our own deepfakes were detectable, and the artefact-level findings from those detectors — in particular an MFCC delta-zero sign reversal in F5-TTS output and a vocoder fingerprint hierarchy in which two-dimensional convolutional vocoders evade detection more effectively than HiFi-GAN or iSTFT vocoders — directly informed our detection-side design.

Throughout, we use *attacker* and *defender* as standard shorthand from the adversarial-robustness literature for the generation and detection roles, respectively.

The asymmetry between our two sub-task scores — **0.9522** on Detection versus **0.4304** on Generation — admits a structural reading under controlled single-team conditions: the defender succeeds when any one of its six backbone encoders captures an artefact, whereas the attacker must simultaneously evade every encoder that any opposing detector might deploy. We refer to this OR-versus-AND geometry as ensemble-level structural asymmetry, and we develop the mechanism in §5.1.

1.3. Contributions

Relative to prior work that augments SSL backbones for cross-corpus robustness [5] or selects intermediate layers within a single SSL backbone for ensemble fusion [6], or that surveys audio-LLM and holistic anti-spoofing strategies [7, 8], this paper contributes the *mechanism behind* the observed phenomena rather than additional points on the leaderboard. Concretely, the paper makes the following five contributions:

1. A four-backbone self-supervised learning ensemble combining WavLM-Large, Wav2Vec2-XLS-R-300M, ECAPA-TDNN, and x-vector representations, with a top-960 conservative threshold submission strategy, achieving a Final Score of **0.9522** on the official Audio Detection evaluation, with perfect accuracy (1.0000) on participant-generated deepfakes and 0.8875 on the held-out organizer ground-truth real data (§4).
2. An Audio Generation system whose official entry — an F5-TTS v1 baseline processed with a uniform reverberation pass, submitted as a deliberate anti-forensic probe — ranked first in the sub-task with a Final Score of **0.4304** (WER 4.99%, CER 2.07%; §3).
3. A cross-track analysis that operationalizes the multi-view detection principle of Singh et al. [9] in the representational geometry of SSL backbones, exposing an ensemble-level structural asymmetry: the defender enjoys an OR-gate advantage (any one of six backbone encoders capturing an artefact suffices), while the attacker faces an AND-gate disadvantage (simultaneous evasion of every encoder required). We support this framework with five generation-side insights (MFCC delta-zero sign reversal, vocoder fingerprint hierarchy, multi-model hybridization, quality–evasion trade-off, and post-processing-induced evasion collapse) transferred to the detection-side design (§5).
4. Four systematically falsified ceiling experiments — layer selection, training-time augmentation, test-time augmentation, and score calibration — for which *none of the four improved the baseline*, followed by a single successful breakthrough, *backbone diversification*, that defines our final submission strategy (§4, §6).
5. A three-region representational geometry across six SSL backbone encoders, with the lowest pairwise correlation we measured ($r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$) attributable to encoder

geometry rather than to dimensionality (PCA-192 ablation: $r = 0.4304$), and an honest *saturation* self-criticism ($r(\text{XLS-R}, \text{HuBERT}) = 0.971$) that distinguishes genuine model behavior from heuristic-driven ceiling effects — collectively serving as empirical evidence for the *architectural-insurance* hypothesis that motivates our multi-backbone design (§6).

1.4. Paper Roadmap

Section 2 surveys related work. Section 3 describes our Audio Generation submission and self-evaluation. Section 4 describes our Audio Detection methodology and official results. Section 5 presents the cross-track analysis and the five generation-to-detection insights. Section 6 reports further experiments on backbone representational geometry, bootstrap confidence intervals, and dimensionality ablation. Section 7 discusses limitations, including the 11.25% false-positive rate on organizer ground-truth real data. Section 8 concludes and outlines perspectives for future work. Because the paper covers both sub-tasks and the analysis that connects them, readers primarily interested in the Detection system may focus on §4 and §6, while §3 and §5 address Generation and cross-track insights respectively.

2. Related Work

Audio deepfake detection has progressed alongside self-supervised speech representations. WavLM [10], HuBERT [11], and Wav2Vec2-XLS-R [12] provide backbones widely adopted in anti-spoofing pipelines, including end-to-end systems such as RawNet2 [13] and AASIST [14], and successive ASVspoof editions [4] have driven cross-dataset generalization studies. Combei et al. [5] and Serrano et al. [6] report SSL-based detectors on public benchmarks; our detection stack (§4) builds on the same encoder family.

A second recent line applies large audio-language models to anti-spoofing. ALLM4ADD [7] reframes detection as an audio-LLM prompting task, and HoliAntiSpoof [8] proposes a holistic multi-source view. Singh et al. [9] formalize a multi-view principle across complementary analysis levels that is directly relevant to the four-backbone architectural insurance we develop in §4.6.

On the generation side, flow-matching text-to-speech (F5-TTS [15]), massively multilingual zero-shot voice cloning (XTTS [16]), and supervised-semantic-token synthesizers (CosyVoice [17]) have raised the bar for zero-shot naturalness, which our official entry exploits through a four-model hybrid (§3.2). Perceptual quality throughout is scored with NISQA [3]. To the best of our knowledge, no directly comparable prior study submits both a defensive and an offensive system to the same edition of ImageCLEF, which motivates the cross-track framing of §5.

3. Audio Generation Task

3.1. Task Setup

The Audio Generation sub-task asks each team to synthesize 480 utterances from 16 target speakers (30 utterances per speaker). Submissions are scored on the product of an audio-quality score (NISQA MOS, word and character error rates, and ECAPA-TDNN / WavLM speaker similarity) and a deepfake-evasion score measured against both participant-built and organizer-built detectors (§3.5). The asymmetry between these two evasion populations — the participant detector family is partly known to us, whereas the organizer detector family is undisclosed — becomes the central limitation of our offensive result and is analyzed mechanistically in §7.3.

3.2. Four-Model TTS Hybrid Strategy

We selected four text-to-speech systems — GLM-TTS, F5-TTS v1 [15], XTTS v2 [16], and CosyVoice 3 [17] — with distinct vocoder families and distributed the 16 target speakers across them, with the intent of forcing any opposing detector to recognize four vocoder fingerprints simultaneously rather than one (Insight C, §5.2). This is the offensive analogue of the architectural-insurance frame developed in §4.6:

just as a defender benefits from encoder diversity, an attacker benefits from generator diversity. Speaker assignments to the four models were determined by per-speaker pilot evaluations (WER on a small held-out set and self-detector AUC), favoring for each speaker the model that achieved the strongest evasion at competitive audio quality. (Our official scored entry — an earlier F5-TTS submission with a uniform reverberation pass — predates this program and is described in §3.5; the program below supplied the self-detectors and insights through which that entry's first-place result is analyzed in §7.3.)

Table 1
Four-model TTS hybrid: speaker assignment and vocoder family.

Model	Vocoder family	Vocoder type	Speakers
GLM-TTS	Vocos2D	2D iSTFT	11 (main)
F5-TTS v1	Vocos	1D iSTFT	3
XTTS v2	HiFi-GAN	GAN	1
CosyVoice 3	internal	flow matching	1
(Bark and Qwen3-TTS attempted, abandoned)			—

We additionally evaluated two systems that did not survive into the final submission: Bark (EnCodec vocoder) was abandoned because its word error rate on our prompts was too high to clear the audio-quality floor, and Qwen3-TTS was abandoned because its generation latency (approximately 15 minutes per sentence on our single-A100 budget) was incompatible with the 480-utterance schedule. Honest reporting of these dead-ends follows the same principle as the four falsified Detection ceilings in §4.5: explicit negative results inform the design decisions we did keep.

3.3. Six Self-Detectors and Generation-Side Evaluation

During the Generation work we built six lightweight detectors over classical hand-crafted feature sets to evaluate our own outputs before submission. These detectors were not deployed in the final Detection submission, which uses self-supervised backbones (§4.2); their purpose was diagnostic — to surface artefact patterns we could iterate against and to provide the quality–evasion measurements used in §5.2.

Table 2
Six self-detectors used for Generation-side evaluation.

Feature	Extraction
mfcc_delta	MFCC 20-dim + delta + delta-delta (mean/std)
lfcc	Linear filterbank 20-dim → power → dB
melspec	Mel spectrogram 40-band (mean/std)
spectral	Centroid / bandwidth / rolloff / flatness / contrast
prosody	F0 (YIN) / RMS / ZCR / voicing ratio
rawstats	Mean / std / max / percentiles / silence ratio

Each feature set was classified with a logistic-regression head and evaluated under leave-one-speaker-out cross-validation on our 16 internal speakers. The diagnostic value of this set was structural rather than absolute: six independent classical detectors produce a per-attack agreement profile, and the mfcc_delta detector in particular yielded the delta-zero sign reversal that informs Insight A (§5.2). We treat this six-detector set as an operational scaffold that generated the cross-track observations of §5, not as a Detection submission candidate.

3.4. Submission Strategy: Cherry-Pick and Temperature

Across approximately 50 generation experiments we observed a monotonic quality–evasion trade-off (Insight D, §5.2): lower-WER configurations tended to be detected more reliably by our self-detectors.

We navigated this trade-off by speaker-level cherry-picking (generating multiple utterance candidates per speaker and selecting the one that best balanced WER against self-detector probability) and by sampling-temperature variation. The five final submissions span this trade-off explicitly.

Table 3
Five candidate Audio Generation submissions and their self-evaluation metrics, ordered by internal self-evaluation ranking (based on Self AUC). This internal ranking is distinct from the official leaderboard placement discussed in §3.5.

Submission	WER	CER	Self AUC	NISQA
v2_cherry	2.56%	0.94%	0.634	3.118
filepick_safe	1.68%	0.64%	0.659	∼3.1
rv2_diverse	3.02%	1.25%	0.820	∼3.1
v2_cherry_temp08	2.12%	—	0.649	—
filepick_temp08	1.64%	—	0.663	—

3.5. Official Evaluation Results

Prior to the multi-model program of §3.2, we had submitted an F5-TTS v1 baseline processed with a uniform reverberation pass (pedalboard[1]; `room_size` $= 0.15$, `wet_level` $= 0.08$; single pass over all 480 utterances) in March, as a deliberate anti-forensic probe: room acoustics were intended to mask vocoder artefacts while costing little quality (self-measured NISQA drop of 0.12). No cherry-picking was applied, preserving per-file stochastic variance (seed randomized per utterance). Self-measured WER 4.96%, CER 2.06%, and NISQA 2.9729 later matched the official measurements (4.99%, 2.07%, 2.9729 — the NISQA figure to four decimal places), validating our evaluation pipeline end to end. Following the organizers' per-user scoring consolidation, the team designated this entry as its single official submission; it ranked first in the Audio Generation sub-task with a Final Score of 0.4304.

The organizer-side evaluation, released on 2026-05-15, produced the decomposition in Table 4 for our submission. The Generation Final Score combines audio quality and evasion as

$$S_{\text{gen}} = Q_{\text{audio}} \times \big(0.7\, E_{\text{part}} + 0.3\, E_{\text{org}}\big), \tag{1}$$

where Q_{audio} is the normalized audio-quality score and E_{part} and E_{org} are the evasion rates against the participant and organizer detector pools; for our official entry, $0.7190 \times (0.7 \cdot 0.6142 + 0.3 \cdot 0.5618) \approx 0.4304$ (Table 4).

Table 4
Official evaluation of the team's two scored Generation entries. The reverberation entry was designated the official team submission (per-team consolidation) and ranked first.

Metric	Official entry (reverb, 1st)	File-picked variant (2nd)
NISQA MOS	2.9729	3.128
Speaker similarity (ECAPA-TDNN)	0.6148	0.5875
Speaker similarity (WavLM)	0.9341	0.9129
Word error rate	4.99%	1.68%
Character error rate	2.07%	0.64%
Audio quality score	0.7190	0.7344
Evasion vs participant detectors	0.6142	0.5690
Evasion vs organizer detectors	0.5618	0.3812
Final score	**0.4304 (1st)**	0.3765 (2nd)

[1] https://github.com/spotify/pedalboard

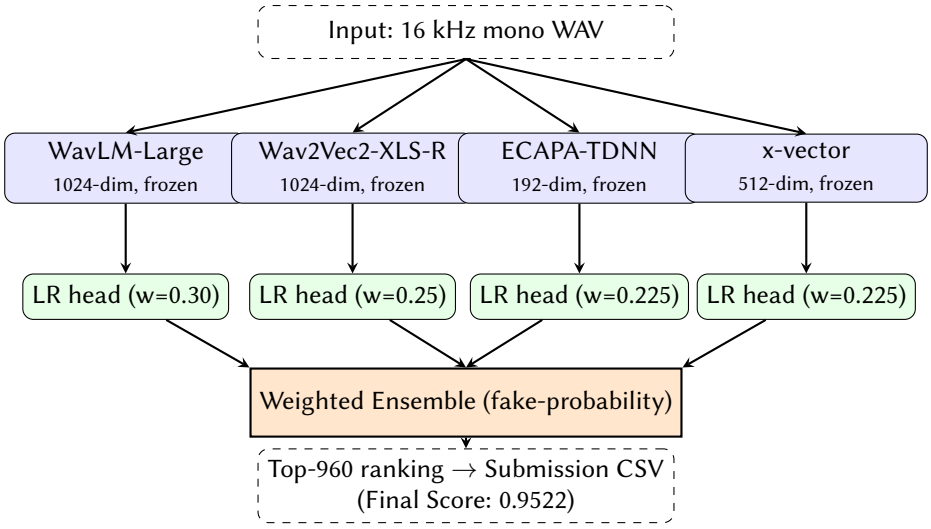

Figure 1: Four-backbone SSL ensemble system for Audio Detection. Each frozen backbone produces utterance-level embeddings independently; per-backbone logistic-regression classifiers are combined via weighted ensemble ($w = 0.30,\ 0.25,\ 0.225,\ 0.225$) to produce the final fake-probability. The top-960 ranking strategy (§4.6.3) maps the ranked probabilities to a binary CSV submission.

Table 4 juxtaposes the two entries. The official reverberation entry evades both detector pools at comparable rates (0.6142 and 0.5618), whereas the file-picked variant shows a wider 0.5690-versus-0.3812 asymmetry; we analyze the sources of this asymmetry — and the internal-vs-official detector-calibration gap that explains the file-picked entry's weaker organizer score — in §7.3.

4. Audio Detection Task

4.1. Task Setup

The Audio Detection sub-task of the ImageCLEF 2026 Deepfake Task [1] requires a binary real-versus-fake decision for each utterance in a test set of $11{,}520$ WAV files. The reference real recordings cover sixteen English speakers and are distributed as 16-bit mono WAV at a 16 kHz sampling rate. Our internal training set comprised 480 real utterances from the Generation-track reference set and $2{,}400$ fake utterances that we synthesized during our own Generation work (480 files for each of five generator variants), giving a 480:$2{,}400$ real-to-fake split before any external extension. The official Final Score is a weighted average of four evaluation categories — baseline deepfakes (weight 0.1), the organizers' real data (weight 0.1), the organizers' held-out real ground-truth data (weight 0.4), and participant-generated deepfakes, weighted (weight 0.4) — so the held-out real and the cross-team fake partitions jointly account for the majority of the score:

$$S_{\text{det}} = 0.1\, A_{\text{base}} + 0.1\, A_{\text{org,real}} + 0.4\, A_{\text{org,GT}} + 0.4\, A_{\text{part,fake}}, \tag{2}$$

where each A term is the accuracy on the corresponding category (baseline deepfakes, organizer real data, organizer held-out ground-truth real data, and participant-generated fake data, respectively).

4.2. Frozen SSL Feature Extraction

Before detailing the feature pipeline, we note that our detection design deliberately commits to frozen self-supervised features and lightweight per-backbone classifiers rather than to end-to-end raw-waveform anti-spoofing architectures such as RawNet2 [13] or AASIST [14]. End-to-end models jointly optimize feature learning and the decision boundary, which can be competitive on single-system benchmarks but entangles representational diversity with optimization noise. By freezing each backbone and fitting only a logistic-regression head on top, we obtain inter-backbone correlations that

reflect each backbone's pretraining geometry rather than co-adapted optimization artefacts, which is what makes the three-region analysis of §6.2 and the architectural-insurance frame of §4.6 interpretable. This design choice is further motivated by our training-set scale: with only 480 real and 2,400 fake utterances available from the Generation-track reference set (§4.1), end-to-end fine-tuning of hundreds of millions of feature-extractor parameters would carry a substantial overfitting risk, whereas frozen SSL features paired with lightweight linear classifiers act as a regularization mechanism for cross-corpus generalization, consistent with recent observations on out-of-domain robustness of SSL-based pipelines [5, 6].

Our final system extracted utterance-level embeddings from four frozen backbones. WavLM-Large (`microsoft/wavlm-large`, approximately 300 M parameters) was run with hidden-state output enabled; we averaged the last four transformer layers (layers 21–24) and then applied a temporal mean-pool, producing a 1024-dimensional vector per utterance. Wav2Vec2-XLS-R-300M (`facebook/wav2vec2-xls-r-300m`) was processed through an identical last-four-mean and temporal-mean pipeline, also yielding a 1024-dimensional vector; matching the two SSL dimensionalities simplified ensembling and side-by-side analysis. Two speaker-recognition encoders completed the set: ECAPA-TDNN (`speechbrain/spkrec-ecapa-voxceleb`, 192-dimensional) and an x-vector encoder (`speechbrain/spkrec-xvect-voxceleb`, 512-dimensional). Every backbone was used strictly as a frozen feature extractor; we performed no fine-tuning, and all inputs were resampled to 16 kHz and downmixed to mono before extraction. Figure 1 summarizes the four-backbone pipeline.

4.3. Classifier and Ensemble

Each backbone fed an independent logistic-regression classifier with standardized inputs ($C = 1.0$, `max_iter` $= 2000$, fixed random seed). We kept the classifier linear by design: with only a few thousand labeled training samples the speaker identities are easily memorized, so a deeper head would be at high risk of overfitting rather than learning a generalizable artefact boundary. The final detector, which we denote v18d, is a weighted average of the four per-backbone probabilities, with weights 0.30 (WavLM), 0.25 (XLS-R), 0.225 (ECAPA-TDNN), and 0.225 (x-vector). The design rationale, developed across the experiments of §4.6 and quantified in §6, is that ensemble robustness against evasive hybrid attacks follows from representational diversity across backbones rather than from backbone count alone; we refer to this property as *architectural insurance*.

We adopt Leave-One-Speaker-Out (LOSO) cross-validation as the primary internal evaluation protocol: for each fold, all utterances from one speaker are held out as the test set while the remaining speakers form the training set, and we report the mean per-speaker area under the ROC curve (AUC) across all folds. Our base configuration uses a 16-speaker pool drawn from the organizer-provided Real recordings; §6 reports a 56-speaker extension that integrates an external Real source.

4.4. Test-Time Sample-Rate Heuristic

The sampling-rate distribution of the test set is markedly non-uniform: 10,080 files (87.5%) are 16 kHz, 480 (4.2%) are 22.05 kHz, and 960 (8.3%) are 24 kHz. Because the reference real recordings and our internal fakes are both exclusively 16 kHz, any non-16 kHz file is, with high probability, a submission from another team that did not resample to the reference rate. We therefore applied a hard rule at submission time: every file with a sampling rate in $\{22{,}050,\ 24{,}000\}$ is labeled fake, which affects 1,440 files (12.5% of the test set). This heuristic was motivated by a Generation-side observation that different vocoders leave distinct resampling spectra, an artefact discussed in §5. Importantly, the rule was applied *only* when producing the submission file; it never entered the leave-one-speaker-out AUC measurements reported in §6, so the reported ceilings reflect genuine model behavior rather than the trivial detection of resampled fakes.

4.5. What Did Not Work — Four Falsified Ceilings

Before the successful experiment described in §4.6, we tested four candidate improvement directions on top of a frozen WavLM + logistic-regression baseline (mean leave-one-speaker-out AUC 0.9985): transformer-layer selection, training-time augmentation, test-time augmentation, and score calibration. None of the four improved the baseline. We report them because consistently documenting what did not work is a direct response to the organizers' request for both positive and negative results, and because the pattern itself is informative: it is consistent with the self-supervised pretraining already supplying a representation robust to the perturbations these interventions introduce.

(C1) WavLM layer selection. Despite prior reports of intermediate-layer sensitivity to spoofing artefacts [6], our last-four baseline (0.9985) matched or exceeded every variant (all-layer 0.9983, layer-12 0.9978, mid-6–12 0.9950, layer-6 0.9784, early-1–6 0.9663) — likely a vocoder-distribution mismatch (HiFi-GAN/MelGAN in prior work vs. Vocos/Vocos2D here).

(C2) Training-time augmentation. Five augmentation schemes (RawBoost, three-second random crop, μ-law codec simulation, additive noise at 15–30 dB SNR, and a combined variant) all stayed within the ± 0.001 LOSO uncertainty band, consistent with the asymmetric-augmentation finding of Combei et al. [5]; in our same-distribution evaluation, WavLM pretraining appears to already cover the relevant noise and codec variation.

(C3) Test-time augmentation. Averaging predictions over three-second crops degraded the score distribution: high-confidence real predictions fell from 1,113 to 486, while the ambiguous middle band grew from 297 to 903. WavLM's bimodal low-probability cluster depends on full-clip semantic context, so full-clip inference was retained.

(C4) Score calibration. Length-shrink rescaling, temperature scaling ($T \in \{0.5, 0.7, 1.0, 1.5, 2.0\}$), and a length-shrink/classical hybrid all degraded the result, with the length-shrink hard cap collapsing the short-utterance distribution. Without held-out ground-truth calibration data, blind post-hoc calibration was a gamble rather than a dependable improvement.

None of the four interventions exceeded the baseline, which leaves the single successful direction analyzed next — backbone diversification (§4.6) — as the only ceiling we were able to raise.

4.6. What Worked — Backbone Diversification

The fifth experiment, and the only one that exceeded the baseline, was backbone diversification. Pairing WavLM with Wav2Vec2-XLS-R-300M last-four features at a 0.6:0.4 probability weighting — the configuration we denote v7b — raised the mean leave-one-speaker-out AUC to 0.9992. Taken alone this is a small absolute gain over the 0.9985 baseline, and we do not interpret the marginal AUC as the meaningful signal. The informative quantity is the per-attack agreement between the two backbones.

4.6.1. The architectural-insurance frame

Per-fold Pearson correlation between WavLM and XLS-R probabilities fell from $r \approx 0.92$–0.97 on the four easier single-vocoder attacks to $r = 0.857$ on the hardest multi-model hybrid — the regime where an ensemble is expected to contribute most. We read this as the two encoders making correlated decisions where detection is easy and less correlated decisions under active evasion, so the ensemble acts as insurance against hybrid attacks rather than as a generic averaging gain. We refer to this property as *architectural insurance*; the full six-encoder representational geometry that quantifies it is developed in §6.

A per-file agreement analysis on 10,080 16 kHz test utterances confirms this attack-conditional structure quantitatively: WavLM and XLS-R disagree on only 148/10,080 files (1.47%), and this small

disagreement set is predominantly concentrated on the multi-model hybrid attack rather than uniformly distributed across attack families. The ensemble's value therefore lies in those 148 files, not in averaging behavior across the easy majority — a mechanistic interpretation we make precise in §6.

4.6.2. From v7b to v18d — adding ECAPA-TDNN and x-vector

We then tested whether speaker-recognition encoders contribute a representation that the two SSL backbones do not. ECAPA-TDNN (192-dimensional, AAM-Softmax-trained on VoxCeleb) sat at $r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$ on the hardest hybrid attack, the lowest pairwise correlation we measured anywhere in this work; it therefore enters the geometry as an isolated region rather than as a near-duplicate of either SSL backbone. The x-vector encoder (512-dimensional, softmax-trained on VoxCeleb) contributed through a different route: its standalone hybrid AUC at the harder 56-speaker setting was 0.9982, well above WavLM's 0.9619 in the same regime, because it did not collapse under the LibriSpeech distribution shift. Combining all four backbones — WavLM, XLS-R, ECAPA-TDNN, and x-vector at weights 0.30, 0.25, 0.225, and 0.225 respectively, the configuration we denote v18d — reached a 56-speaker hybrid AUC of 0.9988, an improvement of +0.0168 over v7b in the same regime. A five-seed, 80%-stratified paired bootstrap places this improvement over v7b in a wholly positive 95% confidence interval, so it is not a single-seed artefact. The detailed geometry and bootstrap analysis are reported in §6. Although evaluating four backbones increases the computational footprint at inference time, all feature extractors are frozen and the per-backbone classifiers are lightweight logistic-regression heads, so the v18d ensemble remains practical to run on a single GPU.

4.6.3. Submission strategy — top-960 conservative threshold

The submission required a binary real-or-fake label for every test file. Rather than thresholding at a fixed probability, we ranked all test files by ensemble fake-probability and assigned the lowest-ranked 960 files to real and the remainder to fake. This top-960 conservative threshold, motivated by the expected scale of competitor fake submissions relative to the test set, was the primary v18d submission rule, and the same ranking-based rule was applied to the v7b backup submission. The §4.4 sample-rate rule was layered on top of this decision at submission time only.

We note that this top-960 strategy is transductive rather than inductive: the cutoff is set by ranking test files, which uses test-time information about the relative score distribution. The binary CSV submission format (0 = real / 1 = fake) does not admit a continuous probability output, so a threshold of some form is required regardless; we chose a ranking-based threshold over a fixed-probability threshold (e.g., 0.5) because the latter is sensitive to the absolute calibration of an uncalibrated ensemble (§4.5, C4), whereas the former depends only on the relative order of predictions. We acknowledge this is a transductive choice and report it as a property of the submission format rather than as a classifier-internal capability.

4.7. Official Evaluation Results

The official evaluation, conducted by the ImageCLEF 2026 Deepfake Task organizers [1], was released on 2026-05-15. Table 5 reports the four scored categories and their weighted contributions to the Final Score.

Table 5
Audio Detection official evaluation results. The Final Score is the weighted sum of the four category contributions.

Category	Score	Weight	Contribution
Baseline Deepfakes	0.9719	0.1	0.0972
Organizers Real Data	1.0000	0.1	0.1000
Organizers Real GT Data	0.8875	0.4	0.3550
Participant Deepfake Weighted	1.0000	0.4	0.4000
Final Score		**0.9522**	

The score of 1.0000 on participant-generated deepfakes (weight 0.4) indicates that the v18d ensemble generalized to generator families it had not been trained against, which is the central robustness claim of this work; cross-generator generalization is exactly the failure mode that the architectural-insurance design of §4.6 was intended to address. The score of 1.0000 on the organizers' public real data is consistent with the top-960 conservative threshold producing no false positives on that partition. The lower score of 0.8875 on the held-out organizer ground-truth real data corresponds to an 11.25% false-positive rate on a previously unseen real distribution; we defer a mechanistic discussion of this gap, and of its possible relationship to the 12.5% non-16 kHz fraction of the test set, to §7.

Note. All leave-one-speaker-out AUC values reported in this paper are computed *without* the §4.4 sample-rate rule; that rule was applied only at submission time. This separation ensures that the reported ceiling values reflect genuine model behavior rather than the trivial detection of resampled fakes. Table 6 summarizes, category by category, which quantities are affected by the rule and which cannot be recomputed without organizer ground-truth files.

Table 6
Effect of the sample-rate rule (§4.4) on the official Detection categories. The rule was applied at submission time; the without-rule counterfactuals require organizer ground-truth files and cannot be recomputed from the public release.

Category (weight)	With rule (official)	Without rule
Baseline deepfakes (0.1)	0.9719	— (GT-dependent)
Organizer real (0.1)	1.0000	— (GT-dependent)
Organizer real GT (0.4)	0.8875	— (GT-dependent)
Participant fake, weighted (0.4)	1.0000	— (GT-dependent)
Final Score	**0.9522**	— (GT-dependent)

The without-rule counterfactuals require the organizer ground-truth partition, which is not publicly released, so the right column is left unspecified. The rule-free measurements that we *can* compute are the internal leave-one-speaker-out AUCs on our own fake variants, reported throughout §4.6 and §6, where the sample-rate rule is deliberately not applied.

5. Cross-Track Insights

5.1. Cross-Track Asymmetry

We submitted to both the Detection and Generation sub-tasks under one single-researcher compute budget (a single A100 GPU), a setting most participating teams did not adopt: a team that enters only one track cannot measure the other side of the same system. The analysis that follows therefore builds on the generation system described in §3 and connects its offensive design choices to the detection-side observations reported below. On the defensive side, our Detection submission reached a Final Score of 0.9522 and identified all (1.0000) participant-generated deepfakes. On the offensive side, our official Generation submission reached a Final Score of 0.4304, ranking first in the Audio Generation sub-task and evading both detector pools at comparable rates (participant-built 0.6142, organizer-built 0.5618); §7.3 discusses why the official entry's comparable rates diverge from the exploratory file-picked variant (0.5690/0.3812). Read together, these results operationalize the multi-view detection principle in the representational geometry of an SSL backbone ensemble. Singh et al. [9] establish that complementary analysis levels (semantic, structural, and signal) improve robustness against advanced text-to-speech families. We specify a measurable form of that principle within a six-encoder ensemble and find that, under our controlled single-team conditions, the relationship between defenders and generators is not symmetric. The defender succeeds when any one of six backbone encoders captures an artefact—regions R_1 (WavLM), R_2 (the content-acoustic cluster), and R_3 (ECAPA-TDNN) each impose an independent

constraint—whereas the attacker must simultaneously evade every encoder that any opposing detector might deploy. The measured $r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$ implies that evading every region at once is substantially harder than the marginal evasion rate against any single detector would suggest (Figure 2). The 0.9522-versus-0.4304 gap is therefore not a comment on the relative quality of our two systems — our official Generation entry traded a modest quality cost (WER 4.99%, CER 2.07%; §3.5) for the strongest evasion in the sub-task, ranking first with a Final Score of 0.4304 — but an empirical estimate of the asymmetry intrinsic to multi-backbone defense within the acoustic-only representational space we sampled (§8 discusses extensions to audio-LLM views) in the current text-to-speech landscape; the structural mechanism is developed in §4.6 (the architectural-insurance frame) and quantified in §6.2 (the three-region representational geometry). We emphasize that this estimate is bounded by an $N = 1$ team setting; multi-team replication is the natural next step for the architectural-insurance frame to be tested against independently-built attacker populations.

This cross-track view is reinforced by the Phase 3 complementarity analysis of §6.1: under the v8 distribution shift the per-fold Pearson correlation between WavLM and XLS-R drops from 0.857 to 0.763, with the response concentrated in the same hardest-hybrid regime that drives the cross-track asymmetry. The defender's disagreement channel and the attacker's evasion gap are therefore not independent phenomena but two facets of the same six-encoder representational geometry. The same mechanism that makes architectural insurance work for Detection (§4.6) also constrains the Generation side's evasion budget — the attacker cannot move all six encoders simultaneously without sacrificing the vocoder-fingerprint coherence that audio quality requires (Insight D, §5.2).

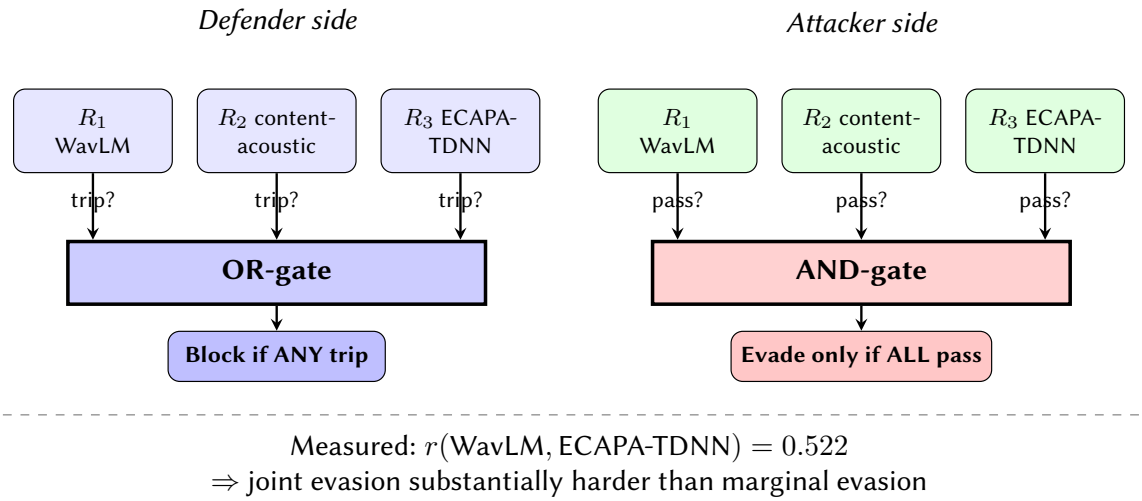

Figure 2: Structural asymmetry between defender and attacker in a multi-backbone SSL ensemble. The defender succeeds when any single region (R_1 WavLM, R_2 content-acoustic, R_3 ECAPA-TDNN) detects an artefact (OR-gate), whereas the attacker must simultaneously evade every region (AND-gate). The measured inter-region correlation $r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$ (§5.1) implies that joint evasion is substantially harder than marginal evasion against any single detector.

5.2. Five Generation-to-Detection Insights

During the Generation work we built six internal detectors and observed five artefact patterns that directly informed the Detection-side design. We summarize each below with its empirical signature and its design implication; the operational consequences are incorporated into the v18d submission strategy (§4.6).

5.2.1. Insight A — MFCC delta-zero sign reversal in F5-TTS

The zero-th MFCC delta coefficient reverses sign between real and F5-TTS audio: real recordings produce a mean delta-zero of approximately -0.47, whereas F5-TTS produces approximately $+1.06$ —

an opposite sign with roughly twice the magnitude. We verified this only on F5-TTS in the present work; whether it generalizes to other vocoder families is not yet established, so we report it as a preliminary observation rather than a universal artefact. The very brittleness of this single-vocoder classical artefact is informative in its own right: it motivates our shift away from hand-crafted features and toward the multi-backbone SSL ensemble in the final detection design, where representational diversity (§6.2) provides the robustness that any single classical detector cannot.

5.2.2. Insight B — vocoder fingerprint hierarchy

Across four vocoder families on the hardest hybrid attack, the self-detector AUC stratifies cleanly by vocoder architecture: HiFi-GAN (XTTS v2) 0.976, the internal CosyVoice 3 vocoder 0.961, Vocos (F5-TTS v1) 0.882, and Vocos2D (GLM-TTS) 0.760. The ordering "GAN $>$ iSTFT $>$ two-dimensional convolution" was consistent across our six self-detectors, with two-dimensional convolutional vocoders the hardest to detect. The design implication is that a Detection ensemble trained primarily on GAN-based fakes is structurally vulnerable to non-GAN vocoders — exactly the scenario that motivated multi-backbone diversification and the top-960 conservative threshold submission strategy (§4.6).

5.2.3. Insight C — multi-model hybridization as the strongest evasion

Single-model fakes were detected at a weighted-mean self-detector AUC of 0.781; distributing speakers across a four-model hybrid (GLM-TTS, F5-TTS v1, XTTS v2, CosyVoice 3) dropped this to 0.634 — a non-linear evasion gain of 0.147. The multi-modal fake-feature space collapses the decision boundary of single-model detectors. This is the offensive analogue of architectural insurance: encoder diversity for defenders, generator diversity for attackers.

5.2.4. Insight D — quality–evasion trade-off

Lower word error rate (higher audio quality) correlated monotonically with higher self-detector AUC (easier detection): configurations near WER 1.64% were detected at AUC ≈ 0.66, while configurations near WER 3.02% reached AUC ≈ 0.82. We attribute this to sampling temperature — lower temperature reduces output variance, sharpens the vocoder fingerprint, and raises detection AUC. A higher-quality opponent therefore leaves a cleaner fingerprint, partially offsetting its quality gain from a detector's perspective.

5.2.5. Insight E — post-processing eliminates evasion

The most counter-intuitive finding is that post-processing helps the detector rather than the attacker: applying a median filter to a four-model hybrid fake raised the MFCC self-detector AUC from 0.634 to approximately 0.999 on the hardest hybrid distribution. Speed perturbation, resampling, and high-frequency noise injection showed the same direction — each operation we tested monotonically increased AUC. This is the empirical foundation of the sample-rate hard rule (§4.4): non-16 kHz files in the test set are, with high probability, the residual of an opposing team's resampling step, and that residual is detectable. The rule was applied only at submission time; the leave-one-speaker-out AUC measurements of §6 deliberately exclude it so that the reported ceilings reflect genuine model behavior, as noted in §4.7.

5.3. Knowledge Transfer Loop

The five insights above describe one direction of transfer, Generation to Detection, and two operational mechanisms drove that direction in our pipeline. Insight E supplied the empirical evidence for the sample-rate hard rule (§4.4), and Insight B's vocoder hierarchy informed the four-backbone diversification behind v18d. Insight D's quality–evasion trade-off further reinforced this direction: higher-fidelity attackers are pushed toward lower sampling temperatures, which sharpen rather than mask the vocoder

fingerprint, providing additional justification for the frozen-SSL classifier pipeline whose features are sensitive to exactly this class of deterministic artefact.

The decisive operational influence was the realization that *we* resampled during our own Generation work, which made it natural to suspect that competing teams *had not* and that their submissions would therefore carry a detectable resampling residual. This single observation produced the §4.4 submission heuristic that contributed the 1.0000 score on the organizers' public real data in our final Detection submission.

The reverse direction was empirically weaker but conceptually symmetric. Building six internal Detection backbones during the Generation work surfaced the three-region geometry (§6.2) at a smaller scale than the final Detection ensemble. We did not have a controlled mechanism to feed this finding back into Generation — the single-team budget and a fixed Generation submission schedule left no room for it — so the offensive system did not benefit from a closed loop. Closing that loop is one of the future-work directions outlined in §8.

The aggregate value of operating both tracks under one compute budget is asymmetric: material defensive improvements (the sample-rate rule, the four-backbone insight) but no comparable offensive gains. The 0.9522-versus-0.4304 gap of §5.1 therefore reflects the asymmetric maturity of the transfer loop as much as the structural asymmetry itself.

6. Further Experiments

6.1. LOSO 56-Speaker Extension

We extended the v7b evaluation with an external real source, LibriSpeech `dev-clean` [18], contributing $1{,}953$ utterances from 40 previously unseen speakers. Combined with the organizer's 16-speaker reference pool, this enlarged the real speaker pool from 16 to 56 and brought the real-to-fake ratio to a near-balanced $1{:}0.97$ ($2{,}433$ versus $2{,}400$). Because the fake pool covers only the 16 internal speaker identities, the held-out folds remain those 16 internal speakers while the 40 LibriSpeech speakers always sit in the training pool; v8 is therefore a controlled real-side distribution shift with the fake distribution held fixed, which lets us test whether the architectural-insurance frame survives such a shift rather than simply adding evaluation folds.

Under v8 the mean LOSO AUC was 0.9960, against v7b's 0.9992, and the four easier single-vocoder attacks were essentially unchanged; only the hardest multi-model hybrid attack moved, with the ensemble AUC dropping from 0.9966 to 0.9819. This movement was asymmetric across backbones: WavLM fell by 0.033 ($0.9951 \rightarrow 0.9619$) while XLS-R rose by 0.002 ($0.9829 \rightarrow 0.9847$), and the per-fold Pearson correlation between the two encoders shifted from 0.857 to 0.763. We interpret this as confirmation rather than regression: the only AUC that responds to the distribution shift is the same one the ensemble was designed to protect, and the response flows through the disagreement channel between the encoders — an architectural-insurance prediction expressed at the data-distribution level.

6.2. Three-Region Backbone Representational Geometry

The six backbones we evaluated — WavLM [10], Wav2Vec2-XLS-R-300M [12], HuBERT [11], the Whisper encoder [19], ECAPA-TDNN [20], and x-vector [21], the latter two via the SpeechBrain recipes [22] — separate into three regions in their pairwise Pearson correlation structure on the hardest multi-model hybrid attack. We introduce this geometry as a refinement of the architectural-insurance frame: ensemble effectiveness is governed by region diversity rather than by backbone count. The regions did not come from a single measurement; they emerged through three successive pre-registered falsifications (v12, v13, v16; §6.5), each forcing a revision of an initially proposed axis.

6.2.1. Region 1 — WavLM (singleton)

WavLM-Large is the only encoder in our accessible ecosystem pretrained with explicit speaker-disentanglement and denoising objectives. On the hardest hybrid attack its predictions correlate with every large content-trained encoder at $r \approx 0.80$ to 0.81 (XLS-R 0.814, HuBERT 0.802, Whisper 0.812) and with ECAPA-TDNN at only $r = 0.522$, placing WavLM consistently outside the Region 2 cluster.

6.2.2. Region 2 — Content-Acoustic Cluster (XLS-R, HuBERT, Whisper, x-vector)

The four large-corpus encoders converge tightly: internal pairwise Pearson $r \geq 0.93$ on the hardest hybrid, with $r(\text{XLS-R}, \text{HuBERT}) = 0.971$ the highest. We characterize this region as content-acoustic rather than masked-prediction, because Whisper (ASR-trained) and x-vector (softmax-trained on speaker labels) violate the latter naming yet still converge with the masked-prediction encoders — indicating that the operative axis is the representational regime, not the training objective. Naively adding more Region 2 backbones to v7b inflates the saturation cluster monotonically (mean hybrid r: $0.857 \rightarrow 0.891$ across v7b/v10/v11/v12) without widening the disagreement channel that protects the hardest hybrid.

6.2.3. Region 3 — ECAPA-TDNN (singleton)

ECAPA-TDNN sits alone: $r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$ on the hardest hybrid (the lowest pairwise correlation we measured anywhere), and $r \leq 0.66$ to every other encoder. The named axis "speaker-recognition specialist" was tested directly by adding x-vector (§6.5) and refuted: x-vector joined Region 2 at correlations up to 0.950, leaving ECAPA-TDNN's particular combination of an AAM-Softmax loss, an SE-Res2 backbone, and a 192-dimensional metric-learned bottleneck as the as-yet-uncharacterized axis.

The architectural-insurance frame predicts that an ensemble spanning all three regions outperforms region-internal expansions. Our final submission v18d (WavLM, XLS-R last-four, ECAPA-TDNN, and x-vector; §4.6) realizes exactly this prediction, and a paired bootstrap (§6.4) confirms that the gain is real.

6.3. Dimensionality Ablation (PCA Verdict)

A natural counter-hypothesis for Region 3's isolation is dimensionality: ECAPA-TDNN produces 192-dimensional embeddings while every Region 2 encoder (the content-acoustic cluster of XLS-R, HuBERT, Whisper, and x-vector), and WavLM, output 1024-dimensional vectors. One might therefore attribute $r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$ to a dimensionality mismatch rather than to a representational difference. We control for this confounder by projecting every backbone onto a common 192-dimensional space with per-encoder principal component analysis fit fold-wise on each LOSO training pool — a projection that retains at least 0.91 of each backbone's variance, ensuring that no test-fold information enters the PCA basis — and re-measuring the pairwise correlations on the hardest hybrid attack.

Under PCA-192, $r(\text{WavLM}, \text{ECAPA-TDNN})$ *decreases* from the native 0.5217 to 0.4304 (and to 0.4287 under PCA-256), moving away from rather than toward the Region 2 cluster. Every cross-region pair involving ECAPA-TDNN falls under dimensionality equalization, whereas the intra-cluster pair $r(\text{XLS-R}, \text{HuBERT})$ rises from 0.971 to 0.9922: the regions separate further, not less, when measured at common dimensionality. The Region 3 singleton position is therefore reinforced rather than explained away by dimensionality control. We had pre-registered three verdict tiers before running the PCA-equalized measurement: Verdict-α when $r(\text{WavLM}, \text{ECAPA-TDNN})$ stays below 0.70 under both PCA-192 and PCA-256 (Region 3 is representational, not a dimensionality artefact); Verdict-β when r lands in $[0.70, 0.85)$ (mixed evidence); and Verdict-γ when $r \geq 0.85$ in either setting (Region 3 is at least partially a dimensionality artefact and the architectural-insurance frame would require a substantive caveat). The measured values 0.4304 (PCA-192) and 0.4287 (PCA-256) both sit well below the 0.70

floor, making this a Verdict-α (strong-validation) outcome: Region 3 reflects a representational property of ECAPA-TDNN's metric-learning recipe, not an artefact of its embedding size.

6.4. Paired Bootstrap Confidence Intervals

Single-point AUC comparisons are vulnerable to file-level noise, especially in the 16-speaker regime where the hardest hybrid AUC is computed over a small fold population. We therefore verified every primary margin claim with a five-seed, 80%-stratified-file paired bootstrap: for each seed we subsampled 80% of the files per attack family while preserving the original real-to-fake ratio, recomputed both the mean LOSO AUC and the hybrid AUC for the two compared systems on the same subsample, and recorded the per-seed margin, holding the classifier seed fixed so that only the data subsample varied.

For the WavLM + XLS-R-last-four + ECAPA-TDNN configuration we denote v15*c, measured against v7b at 56 speakers, the 95% confidence interval for the hybrid-AUC margin is $[+0.0052, +0.0226]$ (point estimate $+0.0089$) and for the mean-LOSO-AUC margin is $[+0.0015, +0.0052]$ (point estimate $+0.0021$); every bootstrap seed agreed in sign on both metrics, so the v15*c improvement is not a single-seed artefact.

For the four-backbone v18d configuration (§4.6) measured against v15*c, a second paired bootstrap places the 56-speaker hybrid-AUC margin in $[+0.0004, +0.0116]$ (individual envelope) and $[+0.0035, +0.0085]$ (standard-error-of-mean envelope), with the 16-speaker margin in $[+0.0000, +0.0006]$ and $[+0.0002, +0.0004]$ respectively. The architectural-insurance frame therefore admits two complementary pathways under one protocol: a low-pairwise-correlation route (v14 and v15* via ECAPA-TDNN) and a route through a standalone backbone that stays robust under the distribution shift (v17 and v18 via x-vector). v18d combines both, and is the only configuration whose margins clear zero in every cell that defines the comparison.

6.5. Pre-Registered Falsifications

The three-region geometry emerged through three pre-registered falsification cycles, each forcing a revision of an initially proposed representational axis (Table 7).

Table 7
Pre-registered predictions and their outcomes on the hardest multi-model hybrid attack.

Experiment	Pre-registered prediction	Outcome
v12 (Whisper)	Whisper joins WavLM (Region 1)	Refuted ($r = 0.812$)
v12 (Whisper)	Region 2 cluster preserved	Confirmed
v13 (ECAPA-TDNN)	ECAPA-TDNN joins WavLM (Region 1)	Refuted ($r = 0.522$)
v13 (ECAPA-TDNN)	Region 2 ceilings $r < 0.85$ hold	Confirmed
v16 (x-vector)	x-vector joins ECAPA-TDNN (Region 3)	Refuted ($r = 0.656$)

Three of the five pre-registered predictions were refuted, revising the framework's named axis twice (denoising-versus-masked-prediction $\rightarrow$ speaker-disentanglement $\rightarrow$ ECAPA-TDNN-specific recipe) before reaching the three-region geometry of §6.2. The pattern that survived — a tight Region 2 cluster with WavLM and ECAPA-TDNN as outliers on independent axes — is the most reliably reproducible claim in our analysis, and the explicit refutations are the mechanism that forced the named axes to converge on representational regime rather than training objective.

7. Discussion and Limitations

We emphasize that the following discussion draws on measurements over a single evaluation cycle without access to organizer ground-truth annotations; the interpretations offered here should be read as hypotheses that a multi-team replication or a private-set inspection could sharpen or refute.

7.1. The 11.25% Real GT False-Positive Gap

Of the four official evaluation categories, the organizers' held-out Real GT Data is the only one on which our submission did not achieve a perfect score: 0.8875 (Table 5), corresponding to a false-positive rate of 11.25% on a real distribution to which we had no prior access. The Note in §4.7 already established that this gap is not produced by the §4.4 sample-rate heuristic operating on the leave-one-speaker-out measurements, because the heuristic was applied only at submission time and the ceiling values of §4.6 and §6 are computed without it. The remaining task is to explain the gap in terms of model behavior.

A suggestive numerical coincidence is worth recording. Of the $11{,}520$ test files, $1{,}440$ (12.5%) are sampled at non-16 kHz rates (22.05 or 24 kHz) and would have been forced to the Fake class by the submission-time heuristic regardless of their true label. Our 11.25% false-positive rate on held-out Real GT is within 0.0125 of this 12.5% figure. Without access to the organizer-side breakdown of Real GT sampling rates we cannot directly verify the connection, but the alignment is consistent with a scenario in which a non-trivial fraction of the organizer Real GT files were stored at non-16 kHz rates and were therefore reclassified as Fake by §4.4. We present this only as a plausible mechanism, not a demonstrated one.

If this reading is correct, the 11.25% gap is a quantifiable trade-off introduced by an explicitly conservative heuristic rather than a failure of the underlying ensemble: the same rule that drove the 1.0000 on the organizers' public Real partition and contributed to the 1.0000 on Participant Deepfake Weighted would have lowered the GT Real score in proportion to the non-16 kHz Real fraction of the held-out set. We acknowledge that this remains a hypothesis, and we flag it as the most operationally significant follow-up question for any future deployment of the v18d submission strategy.

7.2. Cross-Corpus Generalization Risk and the Saturation Caveat

Every leave-one-speaker-out AUC reported in this paper is bounded by our team's five-variant Fake distribution. Configurations from v10 through the v18 weight grid all reach a mean LOSO AUC of 1.0000 at 16 speakers (§6). We read these 1.0000 values as the ceiling of the measurement protocol within the in-distribution Fake population, not as generalization claims. The §6.2 saturation diagnostic ($r(\text{XLS-R}, \text{HuBERT}) = 0.971$) indicates where this ceiling lies: the content-acoustic Region 2 cluster agrees so strongly that adding more Region 2 backbones inflates the saturating signal without widening the disagreement channel that protects the hardest hybrid.

The bootstrap confidence intervals of §6.4 characterize data-distribution uncertainty within this in-distribution hybrid population; they do not bound the cross-corpus drop documented by external benchmarks, which is substantial: detectors near 1.0000 on existing benchmarks can collapse on real-world speech [23], in-distribution detectors lose up to 43% accuracy on unseen TTS families [24], and commercial detectors reach equal-error rates as low as 13.50% on freshly generated in-the-wild deepfakes [25]. We acknowledge this risk explicitly: the v18d primary recommendation rests on within-distribution analysis only, and the architectural-insurance frame becomes empirically falsifiable only when v18d is run against in-the-wild distributions (§8). A second caveat is that our pretrained backbones use heterogeneous training corpora (ECAPA-TDNN on VoxCeleb versus WavLM on Mix 94k); future work must disentangle representational geometry from training-distribution bias before the architectural-insurance frame can be generalized beyond the present pre-training conditions.

7.3. Generation Evasion against Asymmetric Detectors

Our team's two Generation entries — the reverberation baseline designated as our official submission (Final Score 0.4304, ranking first) and the file-picked variant that appears throughout §3.2 (Final Score 0.3765) — were both scored by the organizers, and the ordering they received on the leaderboard is the opposite of the ordering our internal detector suite predicted.

The chronology explains the discrepancy. The reverberation baseline was submitted in March as a deliberate anti-forensic probe (§3.5), when no internal detector was yet available for feedback. During April we built a six-feature leave-one-speaker-out detector suite over the four-model TTS hybrid and

used its self-AUC readings to select subsequent submissions. Two observations from that suite led us to dismiss post-processing approaches, reverberation among them, from further consideration: WavLM features are nearly invariant under mild acoustic post-processing, yielding self-AUC values close to 1.000 on reverberated inputs; and the MFCC delta-sign pattern of Insight A (§5.2) persists through room convolution because it encodes a source-vocoder asymmetry that is not linearly recoverable from filtered output. On the strength of these signals we categorically classified the reverberation family as evasion-ineffective and directed our final submission toward cherry-picked hybrid variants (§3.2).

A retrospective evaluation — run only after the official results were released — quantifies the miscalibration.[2] Under the same LOSO protocol, the reverberation entry scores a self-AUC of 0.868, among the easiest of our candidates to detect internally, whereas the file-picked entry scores 0.659, the profile of a stronger evader. The official scores invert this ordering: $0.6142/0.5618$ (reverberation) against $0.5690/0.3812$ (file-picked). Applying reverberation moved the self-AUC of the underlying F5-TTS audio by only -0.012 ($0.880 \rightarrow 0.868$): our feature set was nearly blind to the very transformation that, on one plausible reading, drove the winning evasion.

One interpretation is that our internal detector suite was calibrated to be more robust to post-processing than the deployed detector ecosystem — an instance of optimizing against a misaligned oracle. The detectors used by other participants and by the organizers appear not to share WavLM's invariance under mild acoustic post-processing, so the same transformation that our suite categorically dismissed as ineffective moved the official evasion score by $+0.0452$ against the participant pool and $+0.1806$ against the organizer pool. This is consistent with Insight D (§5.2): cherry-picking removes stochastic variance in exchange for vocoder-fingerprint purity, and the leaderboard result — where a pre-cherry-pick reverberation entry outscores a cherry-picked hybrid — makes that trade-off empirically visible.

We report this finding with two caveats. First, the two entries differ in more than post-processing alone: the file-picked variant selects among four TTS models per speaker, whereas the reverberation entry uses F5-TTS alone, so attributing the evasion gap specifically to reverberation is directional evidence rather than a controlled comparison. Second, our self-AUC measurements were taken in a development environment, and the retrospective figures reported here reproduce the April measurements only to within 0.012 AUC (see footnote). Both caveats notwithstanding, the mismatch between our internal ranking and the leaderboard ranking is itself the finding: a single-team internal detector suite, even one that reaches near-perfect self-AUC on the transformations it was built to catch, is not a reliable substitute for the deployed detector ecosystem.

8. Conclusion

This paper described our team's submission to the ImageCLEF 2026 Audio Deepfake Detection and Generation task. On the Detection sub-task, a four-backbone self-supervised ensemble combining WavLM-Large, Wav2Vec2-XLS-R-300M, ECAPA-TDNN, and x-vector representations, with a top-960 conservative threshold submission strategy, achieved a Final Score of **0.9522**, including perfect accuracy (1.0000) on participant-generated deepfakes and 0.8875 on the held-out organizer ground-truth real data. On the Generation sub-task, our four-model program (GLM-TTS, F5-TTS v1, XTTS v2, CosyVoice 3; §3) developed detector-informed variants, but our team's official submission — an F5-TTS v1 baseline processed with a uniform reverberation pass — ranked first with a Final Score of **0.4304** (WER 4.99%, CER 2.07%). The controlled comparison between the same team's offensive and defensive capabilities revealed an ensemble-level structural asymmetry: the defender succeeds when *any one* of six backbone encoders captures an artefact, whereas the attacker must simultaneously evade *every* encoder that any opposing detector might deploy. This OR-versus-AND geometry extends the multi-view detection

[2] Retrospective self-AUC figures were measured in July 2026, after the official results were released, in a local environment that reproduces the original April development scores to within 0.012 AUC (v2_cherry 0.646 vs. 0.634; filepick_safe 0.668 vs. 0.659; cherry_short 0.880 vs. 0.882). Development-time self-AUC figures elsewhere in this paper are the original April measurements.

principle of Singh et al. [9] beyond individual analysis levels into a measurable structural asymmetry across a six-encoder representational space, supplying empirical support for our architectural-insurance hypothesis. The Detection-side framework that emerged from this experimentation, which we refer to as *architectural insurance*, attributes the ensemble's hybrid-attack robustness not to encoder count but to representational diversity across the three regions of a six-encoder SSL geometry, with the lowest pairwise correlation $r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$ serving as the primary quantitative signal. Four systematically falsified ceiling experiments preceded the only successful breakthrough (backbone diversification), and an honest saturation self-criticism $r(\text{XLS-R}, \text{HuBERT}) = 0.971$ delimits where genuine representational diversity ends and heuristic-driven ceiling effects begin.

We see three perspectives for future work. First, the *Region 3 hypothesis* remains open. The named axis of "speaker-disentanglement specialist" was refuted, since ECAPA-TDNN sits at $r(\text{WavLM}, \text{ECAPA-TDNN}) = 0.522$ rather than joining WavLM in Region 1, and the second-specialist prediction also failed ($r > 0.94$ between x-vector and Region 2). A direct test of whether the AAM-Softmax with a 192-dim bottleneck recipe is what makes ECAPA-TDNN Region 3 — by substituting TitaNet-Large, which shares ECAPA-TDNN's loss and bottleneck but uses a ContextNet body — is the natural follow-up. Second, every encoder we measured is an acoustic-only view; integrating an audio large language model in the line of ALLM4ADD [7] or HoliAntiSpoof [8] is the only direction in our roadmap that adds a representational regime qualitatively different from the three we have characterized. Third, all measurements in this paper are LOSO-bound on team-internal Fake variants; validating the architectural-insurance claim against an in-the-wild evaluation benchmark is the highest-priority follow-up before any operational deployment.

Author contributions

S. Kim conceived the methodology, designed and conducted all experiments, and authored the manuscript. J. Kim provided extensive advisory input and cross-track review throughout the work. J. Woo provided supervisory feedback on the manuscript.

Acknowledgments

This research was supported by the Ministry of Science and ICT (MSIT), Korea and the Institute of Information & Communications Technology Planning & Evaluation (IITP) under AI University (2026-0-00032, 2026).

Declaration on Generative AI

During the preparation of this work, the author(s) used *Claude* (Anthropic) and *NotebookLM* (Google) in order to: Drafting content, Grammar and spelling check, Paraphrase and reword, Improve writing style, Peer review simulation. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] 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.

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

[3] G. Mittag, B. Naderi, A. Chehadi, S. Möller, NISQA: A deep CNN-self-attention model for multi-dimensional speech quality prediction with crowdsourced datasets, in: Proc. Interspeech 2021, 2021.

[4] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K. A. Lee, J. Yamagishi, ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale, in: ASVspoof Workshop, 2024. ArXiv:2408.08739.

[5] D. Combei, A. Stan, D. Oneata, H. Cucu, WavLM model ensemble for audio deepfake detection, in: ASVspoof Workshop, 2024. ArXiv:2408.07414.

[6] P. Serrano, R. Duroselle, F. Angulo, J.-F. Bonastre, O. Boeffard, Improving out-of-domain audio deepfake detection via layer selection and fusion of ssl-based countermeasures, arXiv preprint arXiv:2509.12003 (2025).

[7] H. Gu, J. Yi, C. Wang, J. Tao, Z. Lian, J. He, Y. Ren, Y. Chen, Z. Wen, ALLM4ADD: Unlocking the capabilities of audio large language models for audio deepfake detection, in: ACM Multimedia, 2025. ArXiv:2505.11079.

[8] X. Xu, et al., HoliAntiSpoof: Audio LLM for holistic speech anti-spoofing (2026). ArXiv:2602.04535.

[9] R. Singh, A. Y. Nair, F. Palumbo, F. Barbaro, A. Dyka, L. Rachakonda, Audio deepfake detection in the age of advanced text-to-speech models (2026). ArXiv:2601.20510.

[10] 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, M. Zeng, F. Wei, WavLM: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing (2022). ArXiv:2110.13900.

[11] 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. doi:10.1109/TASLP.2021.3122291, arXiv:2106.07447.

[12] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, M. Auli, XLS-R: Self-supervised cross-lingual speech representation learning at scale, in: Interspeech, 2022. ArXiv:2111.09296.

[13] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, End-to-end anti-spoofing with RawNet2, in: ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6369–6373. ArXiv:2011.01108.

[14] 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: ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6367–6371. ArXiv:2110.01200.

[15] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, X. Chen, F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching, arXiv preprint arXiv:2410.06885 (2024).

[16] E. Casanova, et al., XTTS: A massively multilingual zero-shot text-to-speech model, arXiv preprint arXiv:2406.04904 (2024).

[17] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, et al., CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, arXiv preprint arXiv:2407.05407 (2024).

[18] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, LibriSpeech: An ASR corpus based on public

domain audio books, in: ICASSP, 2015, pp. 5206–5210. doi:10.1109/ICASSP.2015.7178964.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: ICML, 2023, pp. 28492–28518. ArXiv:2212.04356.

[20] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, in: Interspeech, 2020, pp. 3830–3834. doi:10.21437/Interspeech.2020-2650, arXiv:2005.07143.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-Vectors: Robust DNN embeddings for speaker recognition, in: ICASSP, 2018, pp. 5329–5333. doi:10.1109/ICASSP.2018.8461375.

[22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, Y. Bengio, SpeechBrain: A general-purpose speech toolkit (2021). ArXiv:2106.04624.

[23] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, K. Böttinger, Does audio deepfake detection generalize?, in: Interspeech, 2022. ArXiv:2203.16263 — In-the-Wild (ITW) corpus.

[24] C. Gao, M. Postiglione, I. Gortner, S. Kraus, V. S. Subrahmanian, Perturbed public voices (P^2V): A dataset for robust audio deepfake detection (2025). ArXiv:2508.10949.

[25] Z. Yan, Y. Zhao, H. Wang, VoiceWukong: Benchmarking deepfake voice detection, in: USENIX Security Symposium, 2025. ArXiv:2409.06348.