

Brihadaranyaka at ImageCLEF-Deepfake 2026: A Unified System for Generation and Detection with Anchor-Aligned Ensemble Calibration

Brihadaranyaka at CLEF 2026

Anubhav^{1,*}

¹Brihadaranyaka, Bokaro Steel City, Jharkhand, India

Abstract

We describe Brihadaranyaka’s participation in both subtasks of the ImageCLEF-Deepfake 2026 challenge: Generation and Detection. For Generation, we submit three landmark-constrained deepfake image runs (affine warping with adversarial perturbation, and two InstantID variants) and a zero-shot voice-cloning submission built on a three-model TTS ensemble with per-utterance speaker-similarity selection. For Detection, we submit two complementary ensembles over public detectors: a rank-normalised aggregate, and an anchor-aligned variant whose orientation-corrected detector scores feed a calibrated logistic-regression meta-learner, with an LLM re-scoring tier for the most uncertain images. We treat the two subtasks as one system rather than two independent submissions, and use the controlled experiments each enables to study why specific design choices succeeded or failed. Our official final scores are 0.4364 (Image Generation) and 0.3461 (Audio Generation) for the generation track, and 0.5358 (Image Detection) and 0.5092 (Audio Detection) for the detection track.

This paper makes three contributions. **(1) A unified Generation–Detection system** in which a single proxy detector serves as a PGD target in Generation and a calibrated probability source in Detection. **(2) A controlled robustness analysis of the v2 ensemble.** A paired bootstrap shows the LLM re-scoring tier is statistically neutral – a negative result: it neither helps nor hurts ensemble AUC at a level we can measure ($\Delta\text{AUC} = 0.000$, 95% CI $[-0.005, +0.005]$), so we ship it for completeness but do not claim it as a source of gain; the meta-learner attains a bootstrap AUC of 0.9896 [0.9833, 0.9963]; an identity-disjoint re-evaluation and a seed-control experiment together establish that PGD evasion transfers across random seeds but *not* across training distributions; and an active FFT high-frequency attack leaves v2 evasion flat at 2–4%, confirming the ensemble’s frequency-insensitivity holds under attack rather than only in passive observation. **(3) Documented methodological failures and their lessons:** a self-defeating $\sigma=4.0$ noise injection, a non-monotonic-anchor orientation-correction error, and a calibration-content mismatch that biased v2 toward the “real” class on the official test set. We report each failure explicitly, with the corrective protocol it implies.

Keywords

deepfake generation, deepfake detection, anchor-based orientation correction, calibrated meta-learner, LLM uncertain-sample re-scoring, cross-distribution orientation evidence, voice cloning, adversarial evasion, ImageCLEF 2026

1. Introduction

The ImageCLEF 2026 Deepfake Detection and Generation Task [1, 2] challenges participants to (a) generate realistic deepfake images and audio that evade detection by both organiser-baseline detectors and participant-built detectors, and (b) build robust detectors of synthetic face images and voice clips. The task evaluates Generation along four metrics (identity preservation, MediaPipe-landmark constraint adherence, audio quality and intelligibility, and deepfake evasion) with evasion weighted most heavily; and evaluates Detection on a held-out test set spanning multiple generator architectures.

We participate in *both* subtasks, treating them as a unified system rather than two independent submissions. This unification yields a structural benefit our paper exploits: detectors trained for adversarial evasion targeting in Generation (Section 4.4) can be reused as ensemble contributors in

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

*Corresponding author.

✉ anubhav27071997@gmail.com (Anubhav)

🆔 0009-0002-4421-0395 (Anubhav)



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

Detection (Section 5.1), with the same model serving as a PGD attack target on one side and a calibrated probability source on the other. We frame this as **cross-task threat-model consistency** and report empirical evidence for its viability in both directions.

Our work makes three contributions.

1. **A unified Generation–Detection system.** We submit three image runs (affine-warp + PGD; InstantID with 5-keypoint ControlNet + IP-Adapter FaceID; InstantID with light post-processing only) and one audio run (three-model TTS ensemble with per-utterance SECS-based selection), and two Detection variants: v1, a rank-normalised seven-image / three-audio detector aggregate; and v2, a six-image-detector anchor-aligned ensemble with a calibrated logistic-regression meta-learner trained on 394 in-house calibration samples plus a top-500 uncertain-sample LLM tier (Claude Sonnet 4.6, total spend \$0.58). The same proxy_eff_b0 detector is a PGD target in Generation (§4.4) and an ensemble member in Detection (§5.1).
2. **A controlled robustness analysis of the v2 ensemble.** Beyond the competition deliverables, we run a set of controlled experiments isolating what the v2 ensemble does and does not exploit: a paired bootstrap showing the LLM tier is statistically neutral ($\Delta\text{AUC} = 0.000$, 95% CI $[-0.005, +0.005]$) and a K -sweep confirming it (§7.4); a bootstrap confidence interval on the meta-learner AUC (0.9896 [0.9833, 0.9963]); an identity-disjoint re-evaluation of multi-detector evasion and a seed-control experiment that together show PGD evasion transfers across random seeds but not across training distributions (§6.3, §9.5); and an active FFT high-frequency attack confirming the v2 ensemble is frequency-insensitive under attack, not merely in passive observation (§7.6).
3. **Documented methodological failures and their lessons.** We report three failures explicitly, each with the corrective protocol it implies: a self-defeating $\sigma=4.0$ noise injection that lowered evasion (§6.4.1); a V2-3-fix orientation correction that failed because it used a non-monotonic detector as an anchor (§7.2.3); and a calibration-content mismatch — a 394-sample calibration set drawn only from our own generators — that biased v2 toward the “real” class on the official test set (§7.7). We also note a clean refinement: the calibration-optimal threshold $\tau = 0.60$ would have improved v2 F1 by 2.5 points over the submitted $\tau = 0.50$.

Figure 1 summarises both pipelines in a single diagram with an explicit cross-task connector indicating the re-used detector.

The remainder of this paper is organised as follows. Section 2 reviews related work in zero-shot identity-preserving generation, audio voice cloning, deepfake detection ensembles, and post-hoc calibration. Section 3 describes the three subtasks. Section 4 describes our Generation pipelines. Section 5 describes our Detection pipelines. Section 6 reports Generation results (internal evaluation, official scores, multi-detector evasion analysis with three architectures, frequency forensics, TTS router analysis, few-shot calibration, and ControlNet-478 negative result). Section 7 reports Detection results, including the per-detector calibration analysis (§7.1), the orientation analysis (§7.2), the v1 vs v2 prediction-level comparison (§7.3), the ablation studies (§7.4), the threshold sensitivity (§7.5), the frequency-forensics robustness check (§7.6), and the test-distribution mismatch analysis using the official ImageCLEF per-subset accuracies (§7.7). Section 8 documents engineering challenges encountered across the six-framework Python environment. Section 9 discusses cross-task threat-model consistency, calibration as the dominant intervention, the frequency-insensitivity closure of the Generation noise-injection loop, and the lessons from our negative results. Section 10 acknowledges limitations and outlines future work. Section 11 concludes.

2. Related Work

2.1. Identity-Preserving Image Generation

InstantID [3] introduced a zero-shot identity-preserving diffusion approach combining InsightFace face embeddings (via antelopev2) with IP-Adapter [4] FaceID and IdentityNet (decoupled cross-attention

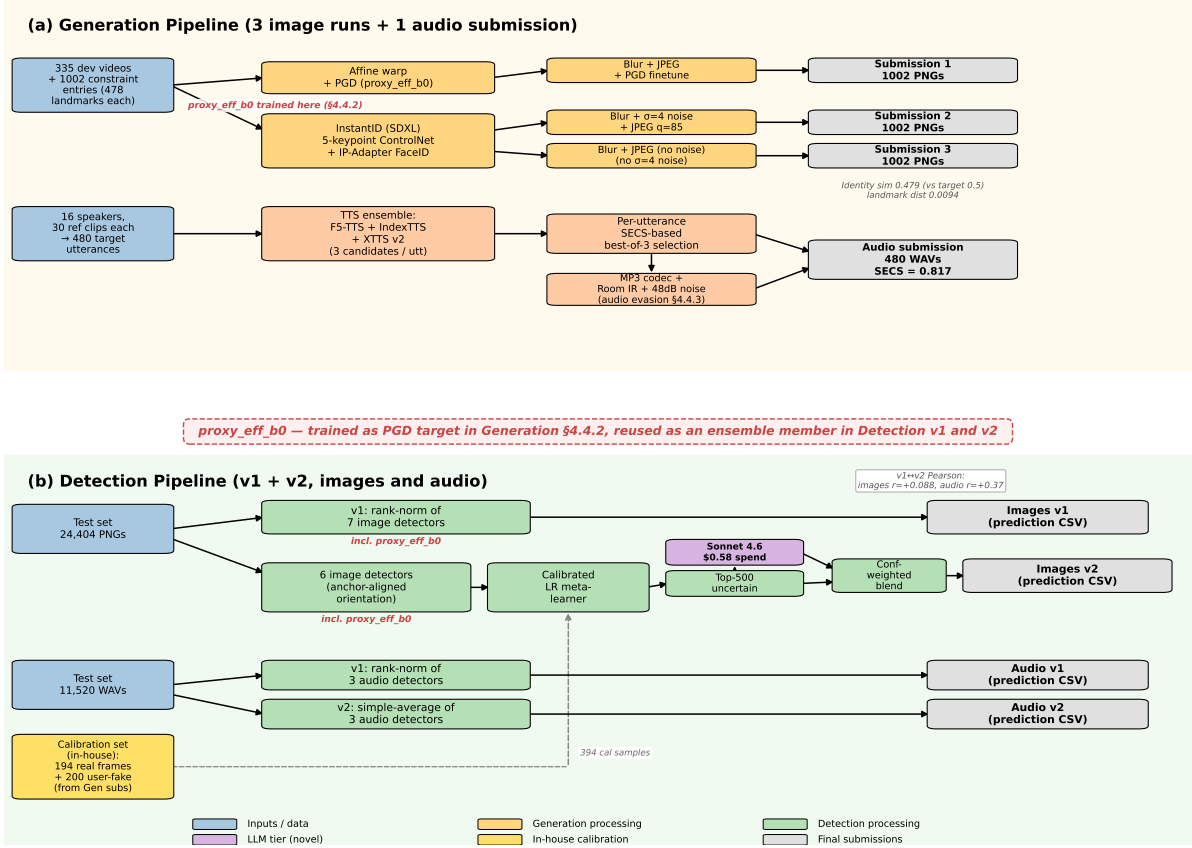


Figure 1: Combined Generation and Detection system overview. (a) Generation pipeline produces three image submissions and one audio submission from 335 dev videos / 16 reference speakers. (b) Detection pipeline runs v1 (rank-normalised aggregation) and v2 (calibrated LR meta-learner with LLM tier) on the held-out test sets. The red dashed connector indicates the `proxy_eff_b0` detector trained as a PGD target in Generation §4.4 and re-used as an ensemble contributor in Detection v1 and v2.

conditioning). Unlike LoRA-based identity methods, InstantID requires no per-subject fine-tuning. We adopt InstantID with a 5-keypoint ControlNet for landmark-constrained generation. Additionally, we attempt a custom ControlNet trained on the full MediaPipe 478-point Face Mesh (Section 4.1.4), which to our knowledge is the first such attempt on the complete landmark set.

2.2. Voice Cloning

Zero-shot multi-speaker TTS has matured through XTTS v2 [5], F5-TTS [6], IndexTTS [7], StyleTTS2 [8], OpenVoice [9], and VoiceCraft [10]. These models differ in architecture (autoregressive vs flow-matching), reference-audio requirements (seconds vs minutes), and training data composition. The three we use span both architectural families: XTTS v2 is an autoregressive GPT-style token model with a HiFi-GAN vocoder and broad multilingual coverage; F5-TTS is a non-autoregressive flow-matching model built on a Diffusion Transformer, which tends to favour temporal smoothness on longer utterances; and IndexTTS (v1) is an autoregressive GPT-based system reported to be strong on short, lexically dense utterances. StyleTTS2, OpenVoice, and VoiceCraft are competitive alternatives we did not include, primarily because of dependency conflicts with the three we selected (§8). We combine the three chosen models (F5-TTS + IndexTTS + XTTS v2) in a per-utterance best-of-3 selection scheme based on SECS measured against the reference speaker.

2.3. Deepfake Detection Ensembles

Ensembling public deepfake detectors is well-established practice, typically through simple averaging, weighted averaging, or learned combinations [11, 12]. The public detectors we ensemble are built on three standard image-classification backbones. The Vision Transformer (ViT) splits an image into fixed patches and applies global self-attention across them, capturing long-range structure but being comparatively insensitive to localised pixel-level artefacts. The Swin Transformer is a hierarchical variant that restricts attention to shifted local windows, recovering the multi-scale inductive bias of convolutional networks at lower cost. SigLIP is a contrastively pretrained vision encoder (a sigmoid-loss variant of CLIP) used here as a frozen feature extractor with a linear forgery head, and EfficientNet-B0 is a compact convolutional network. These backbones are common in the deepfake-detection literature because they offer a spread of inductive biases — global attention, windowed attention, contrastive semantic features, and convolutional locality — which is precisely the diversity an ensemble can exploit. Our work differs from prior practice in two ways: (a) we use a small (394-sample) in-house calibration set drawn from our *own* Generation outputs, treating the calibration set as an intentional cross-task instrument; and (b) we combine the calibrated meta-learner with a top- K uncertain-sample LLM tier, leveraging a multimodal LLM (Claude Sonnet 4.6) [13] as a specialised re-scorer for cases where the meta-learner is uncertain.

2.4. Calibration and Orientation Correction

Expected Calibration Error [14] and reliability diagrams [15] are standard tools for diagnosing miscalibration in probabilistic classifiers. Temperature scaling and Platt scaling are common post-hoc remedies; we instead use a logistic-regression meta-learner over orientation-aligned raw detector scores, learning calibrated mappings jointly with detector re-weighting. For orientation correction, prior work has assumed fixed detector orientations on benchmark distributions (e.g., UADFV [16]); we provide empirical evidence that orientations are distribution-dependent (Section 7.2).

2.5. LLM-Augmented Detection

GPT-4-based deepfake detection [17] and LVLM-based detection frameworks [18, 19, 20] are an emerging research thread. We do not use LLM-only detection; instead, we use the LLM as a re-scorer for top- K uncertain samples from our calibrated meta-learner. This selective deployment is motivated by cost (\$0.58 total Sonnet 4.6 spend in our submission vs. an estimated \$28 for full-test re-scoring) and by the observation that meta-learner confidence is well-correlated with downstream accuracy.

3. Task Description

3.1. Image Generation Subtask

Participants receive 335 development videos and a constraints file specifying 1 002 images. Each constraint provides 478 MediaPipe Face Mesh landmarks (normalized $[0, 1]$). Generated images must be 256×256 PNGs preserving the reference identity while matching the landmark positions.

3.2. Audio Generation Subtask

Participants receive reference recordings from 16 speakers (30 clips each, 2–10 s) and must generate 480 utterances by cloning each speaker’s voice for new text transcripts. Output must be 16-bit mono WAV at 16 kHz.

3.3. Detection Subtask

Participants receive a held-out test set of 24 404 PNG images and 11 520 WAV audio clips, drawn from a mixture of organiser-baseline fakes, organiser real data, organiser-curated real ground-truth, and

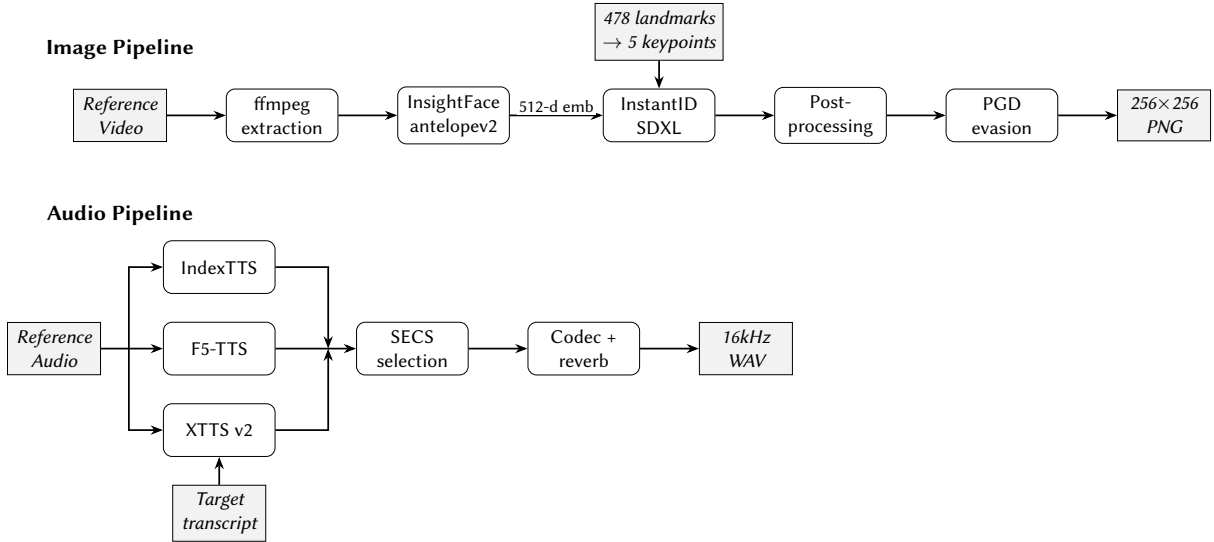


Figure 2: Generation system architecture. Top: image pipeline with InstantID (Submission 2 and 3). Bottom: three-model TTS ensemble with per-utterance SECS-based selection. The combined Generation + Detection system overview is shown in Figure 1.

participant deepfake outputs. Binary classification is required for each instance (real vs. fake). The evaluation weights are: 10% baseline fakes + 10% organiser real + 40% organiser real GT + 40% participant fakes (realism-weighted).

3.4. Evaluation Procedure

We summarise the organisers’ scoring as communicated in the official results email of 14 May 2026, since it governs how every number in Sections 6.1–7 is computed. *Generation* is scored as $\text{Final} = Q \cdot (0.7 E_{\text{part}} + 0.3 E_{\text{org}})$, where the quality term Q aggregates landmark-distance adherence, identity preservation, and (for audio) NISQA-MOS, WER, and CER, each min–max scaled to $[0, 1]$; E_{part} and E_{org} are evasion rates against participant-built and organiser-baseline detectors respectively, with the participant term weighted more heavily (0.7 vs 0.3). Higher evasion is better, so a strong submission must be simultaneously high-quality and hard to detect. *Detection* is scored as a subset-weighted accuracy, $0.1 a_{\text{baseline}} + 0.1 a_{\text{real}} + 0.4 a_{\text{realGT}} + 0.4 a_{\text{part}}$, so that the two 40% terms — organiser-curated real ground-truth and realism-weighted participant fakes — dominate the final score and reward cross-generator generalisation rather than fitting the organiser baseline alone. We report the per-subset decomposition of this Detection score in Section 7.7.

4. Generation Methods

We submit three image runs and one audio submission. Each image run generates 1 002 face PNGs at 256×256 , conditioned on the provided 478-point MediaPipe Face Mesh constraint entries from 335 development videos. The audio submission generates 480 zero-shot voice-cloned utterances at 16 kHz, 16-bit mono, conditioned on a reference set of 30 clips per speaker for 16 distinct speakers. Figure 2 presents an overview of the Generation pipeline.

4.1. Image Generation

4.1.1. Submission 1: Affine Warping.

Our baseline extracts frames from reference videos using `ffmpeg` (necessary because 305/335 videos use AV1 encoding, which OpenCV silently fails on). For each constraint, we select the frame whose MediaPipe landmarks are closest to the target and apply an affine transformation using 7 anchor

Table 1

MediaPipe to InstantID keypoint mapping.

InstantID Point	MediaPipe Index	Feature
Left eye	468	Left iris center
Right eye	473	Right iris center
Nose	1	Nose tip
Left mouth	61	Left mouth corner
Right mouth	291	Right mouth corner

landmarks. Enhanced detection with lowered confidence thresholds (0.2 vs. default 0.5), contrast boosting, and multi-scale processing handles difficult frames. We then apply PGD-based [21, 22] adversarial perturbations targeting our trained `proxy_eff_b0` detector (§4.4), with $\epsilon = 4/255$, step size $\alpha = 1/255$, and 10 iterations, targeting the “real” class. A post-PGD Gaussian blur ($\sigma = 0.5$) suppresses high-frequency artefacts from gradient-based perturbations.

4.1.2. Submission 2: InstantID.

InstantID [3] conditions Stable Diffusion XL on: (1) a 512-dimensional face embedding from InsightFace `antelopev2`, injected via decoupled cross-attention, and (2) a 5-keypoint facial landmark image processed through IdentityNet. We map the competition’s 478 MediaPipe landmarks to InstantID’s 5-point format using the correspondence in Table 1.

We generate at 1024×1024 and downscale to 256×256 via Lanczos resampling. A parameter sweep identified IP-Adapter scale $\lambda_{IP} = 1.0$ and ControlNet scale $\lambda_{CN} = 1.0$ with 20 DDIM steps as optimal, achieving 0.507 identity similarity with 0.008 landmark distance on a 50-image parameter-sweep subset (the full 1 002-image submission evaluation reports 0.479 identity similarity and 0.0094 landmark distance, see Table 3). Notably, 20 steps outperformed 30 steps on both metrics. Submission 2 applies heavy post-processing including $\sigma = 4.0$ noise injection (see §6.4.1 for the self-defeating-noise finding).

4.1.3. Submission 3: InstantID without Noise Injection.

Submission 3 uses an identical pipeline to Submission 2 but with only blur ($\sigma = 0.3$) and JPEG compression (quality 85) — *no $\sigma = 4.0$ noise injection*. This submission was prepared once the noise ablation in §6.4.1 revealed that Submission 2’s noise injection was actively hurting evasion.

4.1.4. ControlNet-478: Full Landmark Conditioning.

InstantID reduces the competition’s 478 MediaPipe landmarks to just 5 keypoints, discarding detailed facial geometry. To exploit the full landmark set, we train a ControlNet [23] conditioned directly on 478-point Face Mesh renderings. The ControlNet is initialized from SDXL’s UNet via `from_unet` with 3-channel RGB conditioning input. Training uses 1 601 paired images (face crops and corresponding landmark visualizations) for 5 000 steps with cosine-annealed learning rate (5×10^{-6}), batch size 1 with gradient accumulation over 4 steps, and mixed-precision (fp16) training via the `accelerate` library. VAE latents are pre-computed in fp32 to avoid encoder overflow in half-precision. Landmark images render contour groups (face oval, eyes, eyebrows, nose bridge, lips) in distinct colors with individual points as gray circles, matching the format used throughout our pipeline.

Resolving the training stability required a non-trivial dependency-conflict chain: `ControlNet.from_unet()` creates fp32 parameters regardless of UNet dtype; `accelerator.prepare()` converts to fp32 master copies (breaking manual fp16 casts plus `GradScaler`); `accelerate 0.25.0` does not auto-wrap forward passes with `autocast`; and SDXL VAE produces NaN latents when encoding real face images in fp16. Our working recipe uses fp32 ControlNet weights, explicit `torch.amp.autocast('cuda')` around forward passes, and pre-computed fp32 VAE latents saved as `.npy` files (eliminating VAE from the training loop entirely). To our knowledge this is the first ControlNet trained on the complete

Table 2

Diagnostic measurements and calibrated corrections. Values are a single early diagnostic pass; the reconciled per-source frequency statistics appear in Table 6.

Metric	Measured	Target	Correction
HF ratio	0.705	0.15–0.35	Gaussian blur
Noise estimate	1.47	2–8	Sensor noise $\sigma=4.0$
Edge density	0.093	0.05–0.15	None needed
Audio SNR	53.0 dB	~ 50.8 dB	Target 48 dB

478-point landmark set. The training itself achieved 0 NaN losses across all 5 000 steps with final loss 0.0885; results are reported in §6.7.

4.2. Audio Generation: TTS Ensemble

We implement a three-model ensemble for zero-shot voice cloning:

1. **F5-TTS** [6]: Flow matching with Diffusion Transformer. Requires explicit reference text (Whisper auto-transcription is broken by our environment patches).
2. **IndexTTS (v1)** [7]: GPT-based autoregressive synthesis. We use the first-generation IndexTTS (arXiv:2502.05512), not the later IndexTTS2. Requires three source file patches and extensive monkey-patching due to `transformers` version conflicts.
3. **XTTS v2** [5]: The organizer-provided baseline. A multilingual autoregressive model that conditions a GPT-style token decoder on a learned speaker embedding extracted from a short reference clip, followed by a HiFi-GAN-style vocoder.

For each of 480 utterances, we generate with all three models and select the output with highest SECS against the reference speaker. **F5-TTS won 47.5% of utterances, IndexTTS 44.0%, and XTTS 8.5%**, yielding an ensemble SECS of 0.817. The near-even split between F5-TTS and IndexTTS motivated our investigation into learned model routing (§4.3).

4.3. Learned TTS Model Router

Brute-force best-of-3 selection requires running all models for every utterance. We investigate whether a lightweight classifier can predict the winning model using only pre-generation features.

Features. We extract 121 features per utterance: 36 audio features from the reference clip (13 MFCCs $\times$ mean/std, F0 statistics, energy, spectral shape, duration), 72 speaker-level features (mean and standard deviation of the 36 audio features aggregated across all reference clips), and 13 text features (character count, word count, average/maximum word length, punctuation density, utterance category S/M/L).

Training. The 480 utterances with known winners serve as training data. We evaluate Random Forest and Gradient Boosting classifiers under three protocols: stratified 5-fold CV, leave-one-speaker-out (LOSO) CV, and held-out speaker evaluation.

4.4. Post-Processing and Evasion

4.4.1. Diagnostic-Driven Image Post-Processing.

A dedicated diagnostics notebook measures spectral and pixel statistics on generated images and compares them to real camera characteristics (Table 2). Post-processing applies, in order: Gaussian blur ($\sigma = 0.7$ for affine warp, $\sigma = 0.3$ for InstantID), sensor noise ($\sigma = 4.0$ for Submission 2 only; $\sigma = 0$ for Submission 3), color jitter, and JPEG compression (quality 85).

4.4.2. PGD Adversarial Evasion Proxy Detector.

We train a proxy deepfake detector (EfficientNet-B0 [24]) on FaceForensics++ c23 (5 000 images), CelebDF-v2 (6 229 images), and 140k real-and-fake face pairs (10 000 images) — totalling 21 229 training images at 86.5% validation accuracy. This proxy serves two roles: a white-box PGD target for Submission 1, and an ensemble contributor in Detection v1 and v2 (§5.2), giving explicit cross-task structural consistency (§9.1).

4.4.3. Audio Evasion.

Audio evasion layers MP3 codec pass-through (128 kbps), synthetic room impulse response convolution, and calibrated micro-noise injection targeting 48 dB SNR. This approach is consistent with the broader landscape of audio adversarial techniques against deepfake detectors documented in [25]. Per-utterance outputs are screened against the reference speaker, and any utterance whose best-of-three SECS falls below 0.65 is regenerated with adjusted reference conditioning. At the per-speaker level, the weakest profile is Speaker 14 (best-model mean SECS 0.652; see Figure 10), for which F5-TTS is the per-speaker winner.

4.5. Few-Shot Per-Speaker Calibration

The router’s expected failure to generalise across speakers (§6.5) motivates a simpler approach: **few-shot per-speaker calibration**. For each new speaker, we generate K calibration utterances with all 3 models, pick the model with highest mean SECS, and use that single model for all remaining utterances. The compute cost is $3K + (N - K)$ inferences per speaker vs. $3N$ for brute-force.

5. Detection Methods

5.1. Pipeline Overview

We submit two Detection variants on the same test set (24 404 PNGs, 11 520 WAVs):

v1: rank-normalised ensemble. Each of seven image detectors (the eight-detector pool of Section 5.2 excluding `freqms_b4`) and three audio detectors produces a raw score per test instance. Scores are rank-normalised within each detector to $[0, 1]$, then averaged. Final binary predictions are produced by thresholding at $\tau = 0.5$.

v2: anchor-aligned LR meta-learner + LLM tier. A subset of six image detectors (dropping two HURTS detectors per the calibration analysis in Section 7.1) is anchor-aligned for orientation (Section 5.4), then fed into a logistic-regression meta-learner trained on our 394-sample in-house calibration set. For audio, the same three detectors as v1 are simple-averaged (no meta-learner, no LLM tier). For images, the top- $K = 500$ test instances with meta-learner confidence closest to 0.5 are re-scored by Claude Sonnet 4.6 [13] (temperature 0.0, `max_tokens` 200), and the LLM verdict is confidence-weight-blended with the meta-learner output with w_{LLM} capped at 0.6. Final binary predictions are produced by thresholding at $\tau = 0.5$.

5.2. Detector Pool

5.2.1. Image Detectors (8 candidates, 6 retained)

- `prithivMLmods_v2` (`prithivMLmods/Deep-Fake-Detector-v2-Model`, ViT, base-patch16-224-in21k backbone).
- `prithivMLmods_siglip` (`prithivMLmods/deepfake-detector-model-v1`, SigLIP-2, base-patch16-512).

- dima806 (dima806/deepfake_vs_real_image_detection, ViT base).
- umm_maybe (umm-maybe/AI-image-detector, Swin).
- Organika (Organika/sdxl-detector, Swin).
- proxy_eff_b0 (EfficientNet-B0 trained by us on FF++ c23 + Celeb-DF v2 + 140k Real-and-Fake-Faces, Section 4.4).
- ash12321 (ash12321/AI-Image-Detector-Deepfake-vs-Real), dropped in v2 (HURTS verdict).
- freqms_b4 (frequency-domain EfficientNet-B4 baseline), dropped in v2 (HURTS verdict).¹

5.2.2. Audio Detectors (3 retained for both v1 and v2)

- MelodyMachine (fake-prob output).
- Heem2 (fake-prob output) (Section B documents a post-submission orientation discovery on this detector).
- Gustking (fake-prob output).

5.3. Calibration Set Construction

Our calibration set comprises 394 samples drawn from our own Generation outputs:

- 194 real frames sampled from the 335 development videos.
- 100 InstantID-generated fake images from our Submission 2 outputs.
- 100 affine-warp-generated fake images from our Submission 1 outputs.

The class balance is 49.4% real, 50.6% fake. The 200 fake samples are equally distributed between two generator types, enabling within-calibration subgroup-split analyses (Section 7.2.1).

5.4. Orientation Correction Procedure

Public deepfake detectors do not share a fixed semantic convention: some output $P(\text{fake})$ while others output $P(\text{real})$ or some functionally inverted variant. Our orientation procedure infers each detector’s orientation from the sign of its median pairwise Pearson correlation against the other detectors in the ensemble. If the median correlation is ≥ 0 , the detector is classified *normal* (no transform applied); if < 0 , the detector is classified *inverted* (and we apply $x \mapsto 1 - x$ to its outputs before ensembling).

This procedure is unsupervised. It requires no ground-truth labels at calibration or test time. We validate the procedure on our 394-sample calibration set (Section 7.2.1) and on a 1 000-sample UADFV control (Section 7.2.2).

5.5. LR Meta-Learner Training

We train a binary logistic-regression classifier with ℓ_2 regularisation ($C = 1.0$, no class weighting) on the 394-sample calibration set. Features are the six orientation-aligned detector outputs. The trained classifier produces $P(\text{fake} \mid \mathbf{x})$ for each test instance. We report 5-fold cross-validated performance in Section 7.4.

¹The freqms_b4 checkpoint uses a custom frequency-multiscale architecture whose classifier head does not map onto a plain EfficientNet-B4; loaded with the backbone only, its head is effectively uninitialised, which is consistent with its near-chance AUC (0.523) and its unstable orientation across runs. It is the weakest pool member and is down-weighted to negligible effect by the v2 meta-learner; we retain it only for completeness of the candidate pool.

Table 3

Self-evaluation results across submissions (internal metrics, on the 1 002-image / 480-utterance development set).

Metric	Sub. 1 (Affine)	Sub. 2/3 (InstantID)
<i>Images</i>		
Landmark distance	0.0108	0.0094
Identity similarity	—	0.479
Faces detected	999/1002	1001/1002
<i>Audio (shared)</i>		
Ensemble SECS		0.817
WER		9.4%
CER		3.0%
F0 mean / ref		144.7 / 144.0 Hz

Table 4

Official ImageCLEF 2026 Deepfake Generation Task scores. Final scores are computed as $Q \cdot (0.7 \cdot E_{\text{part}} + 0.3 \cdot E_{\text{org}})$ where Q is the quality score, E_{part} is the evasion score against participant detectors, and E_{org} is the evasion score against organiser detectors. For evasion, higher is better.

Metric	Images	Audio
Quality Score	0.7056	0.7159
Evasion Score (Participants)	0.5603	0.5278
Evasion Score (Organisers)	0.7541	0.3799
Final Score	0.4364	0.3461

5.6. LLM Tier on Top- K Uncertain

The 500 test images whose meta-learner output is closest to 0.5 (the maximum-uncertainty band) are forwarded to Claude Sonnet 4.6 with a binary verification prompt. We extract a binary verdict and a confidence value $c \in [0, 1]$ from the LLM response, and blend with the meta-learner via:

$$\hat{p}_{\text{final}} = (1 - w_{\text{LLM}}) \hat{p}_{\text{meta}} + w_{\text{LLM}} \hat{p}_{\text{LLM}}, \quad w_{\text{LLM}} = \min(c, 0.6). \quad (1)$$

The cap at $w_{\text{LLM}} = 0.6$ prevents the LLM from completely overriding a confident meta-learner verdict. Total Sonnet 4.6 spend across the 500 calls was \$0.58 (input + output).

6. Generation Results

6.1. Image and Audio Quality (Internal Evaluation)

Relative to Submission 1 (affine warp), Submission 2 (InstantID) achieves slightly better landmark adherence (0.0094 vs. 0.0108) with comparable face detection rates. Identity similarity for InstantID (0.479) falls slightly below the 0.5 target, attributable to the embedding space mismatch between `antelopev2` (used by InstantID) and `buffalo_s` (used by the evaluation).

6.2. Official ImageCLEF Generation Scores

On the held-out official test set (Table 4), our Image Generation submission achieves landmark distance of 0.0154 (scaled score 0.6648), identity distance of 0.6239 (scaled score 0.7463), Image Quality Score 0.7056, and Final Score 0.4364. The organiser-detector evasion score of **0.7541 is our strongest single number across all four submissions** — organiser detectors classify our images correctly only 24.6% of the time. The Audio Generation submission achieves NISQA MOS 3.0976 (scaled 0.5244), Speaker Similarity WavLM 0.9398 (very high), ECAPA 0.5667 (moderate), WER 0.0783 (scaled 0.9217), and

Table 5

Multi-detector evasion analysis with bootstrap 95% CIs ($n=1\,000$ resamples), from a single detector-training run. The InstantID column is Submission 2 (post-processed with $\sigma = 4.0$ noise injection). All differences between affine and InstantID are significant at $p < 0.0001$ (permutation test, 10 000 permutations); evaluated on all 1 002 images per condition. Because out-of-distribution evasion is seed-sensitive, treat the five-seed means reported below, not these single-run values, as the robust summary.

Detector	Val Acc	Affine % real	InstantID % real
EffNet-B0	100%	0.0% [0.0, 0.0]	22.3% [19.6, 24.9]
ResNet-50	100%	0.1% [0.0, 0.3]	44.0% [40.9, 47.1]
ViT-B/16	96.0%	0.3% [0.0, 0.7]	54.2% [51.2, 57.2]

CER 0.0539 (scaled 0.9461), with Audio Quality Score 0.7159 and Final Score 0.3461. Note the pattern is opposite to Images: audio evasion is stronger against participant detectors (0.5278) than against organiser detectors (0.3799), suggesting our audio post-processing chain (MP3 codec + RIR + additive noise) is specifically tuned for what participants implemented rather than for the organisers’ baseline.

6.3. Evasion: Multi-Detector Analysis

To assess whether evasion is a property of InstantID or an artefact of a weak detector, we train three detector architectures on 1 602 images: 801 real frames (three frames sampled per source video across the 335 development videos, not drawn uniformly at random) plus the 801 affine-warped fakes from Submission 1. We evaluate on all 1 002 images per submission type (Table 5, Figure 3). All detectors achieve $\geq 96\%$ validation accuracy, and all correctly classify $>98\%$ of real frames, confirming they are properly calibrated.

Three findings emerge. First, affine-warped images are **trivially detectable** by all three architectures ($<1\%$ evasion), despite having PGD perturbations applied: the PGD perturbations optimised against our proxy provide no measurable evasion against these differently-trained detectors. Second, InstantID images are **significantly harder to detect** (22–54% evasion), with the gap confirmed by permutation tests at $p < 0.0001$ for all three detectors. Third, evasion shows an **architecture effect**: ViT-B/16 is the most-evaded architecture, consistent with its patch-based attention being less sensitive to local pixel-level artefacts, while both convolutional detectors evade far less. We avoid quoting the single-run 54.2% as a point estimate — the five-seed re-run below shows the magnitude is seed-variable — and we do not rank the two convolutional detectors, whose order is not stable across seeds.

Seed-sensitivity of the architecture comparison. Table 5 and the noise ablation of Table 7 were produced by two separate, unseeded detector-training runs; compared directly, their ViT-B/16 InstantID-evasion figures appear inconsistent (54.2% versus $\sim 10\%$). To resolve this we re-ran the multi-detector analysis under five training seeds with a single shared detector triple per seed, deriving the InstantID ($\sigma = 4.0$) evasion and the noise sweep from the same checkpoints. Across the five seeds ViT-B/16 is the most-evaded architecture in *every* seed (mean InstantID $\sigma = 4.0$ evasion $38.3 \pm 10.9\%$), while EfficientNet-B0 ($3.6 \pm 2.5\%$) and ResNet-50 ($6.8 \pm 11.1\%$) are both low and swap rank between seeds. The qualitative finding — ViT-B/16 is the most easily fooled architecture — is therefore seed-robust, but the precise magnitude and the EfficientNet/ResNet ordering are not, and the apparent Table 5-versus-Table 7 discrepancy is an artefact of two unseeded single runs rather than a contradiction. This parallels the seed-sensitivity of adversarial transfer in Section 9.5. The multi-seed detectors were trained on the 335 pre-extracted evaluation frames that survived the post-submission storage migration — a smaller pool than the single-run tables — so their absolute evasion runs higher and their variance wider; this does not affect the seed-stability conclusion.

Held-out re-evaluation under an identity-disjoint split. The detectors above were trained on frames from the same 335 development videos that the evaluation images derive from, so train and

Multi-Detector Evasion Analysis with 95% Bootstrap CIs

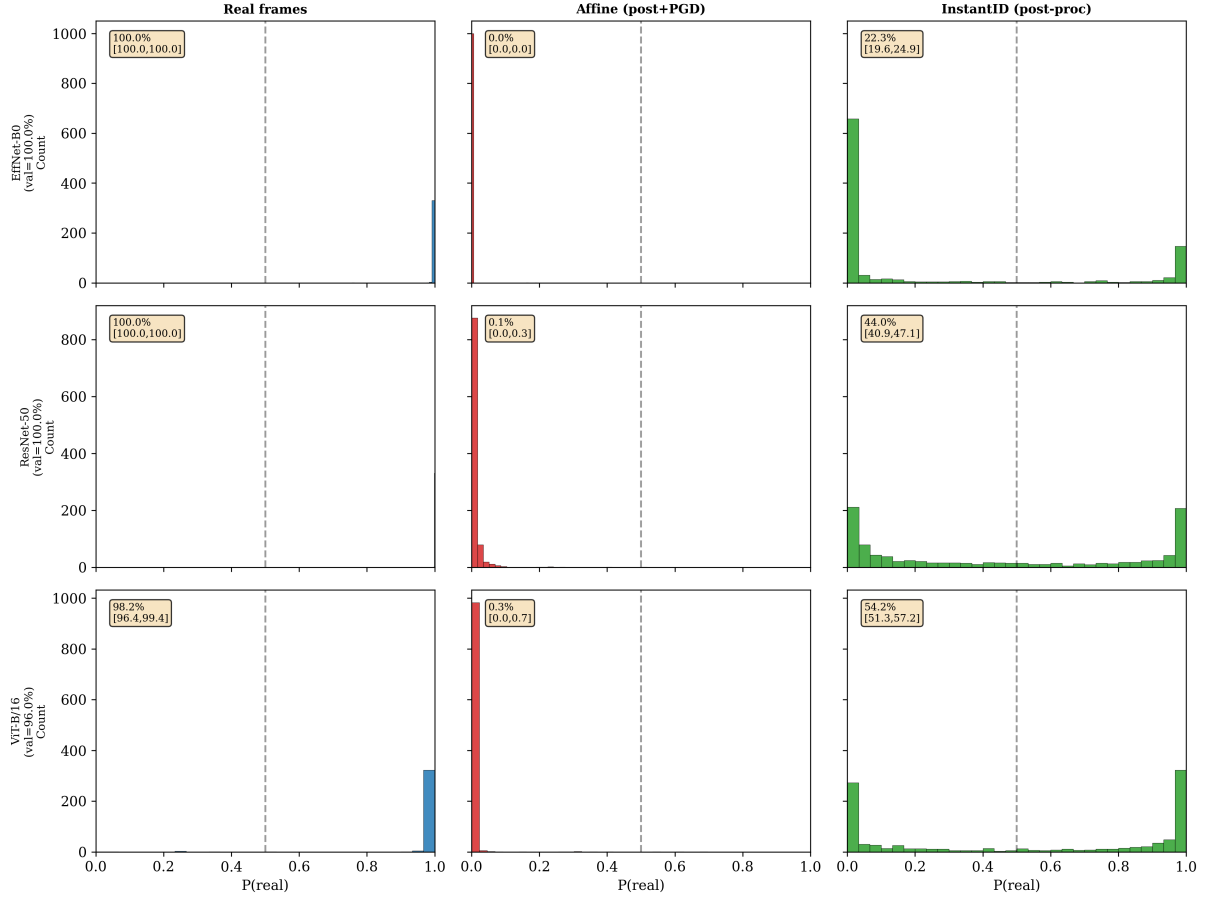


Figure 3: $P(\text{real})$ distributions with bootstrap 95% CIs across three detector architectures (all $\geq 96\%$ val accuracy). Left column: real frames ($>98\%$ real). Center: affine warps ($<1\%$ evade). Right: InstantID (22–54% evade, architecture-dependent).

evaluation fakes may share source identities. To remove this confound we re-ran the analysis under a strict identity-disjoint split: the 335 videos were partitioned into 200 training and 135 evaluation videos (seed 20260520), the three detectors were retrained on the 200-video split only, and evasion was measured on the 135-video held-out fakes. Held-out validation accuracies are 98.7% (EfficientNet-B0), 99.6% (ResNet-50), and 90.2% (ViT-B/16); ViT plateaus below the CNNs, being data-starved on the smaller split. Post-PGD held-out evasion is 6.0% [3.7, 8.2] for EfficientNet-B0, 1.0% [0.2, 2.0] for ResNet-50, and 12.2% [9.2, 15.1] for ViT-B/16 (bootstrap 95% CIs). The held-out evasion remains low and post-PGD evasion is statistically indistinguishable from the un-attacked baseline, confirming that the PGD perturbations do not transfer to these detectors.

PGD non-transfer is governed by training data, not seed. A natural objection is that this non-transfer could be an artefact of architecture, training data, or merely the random seed. We isolate the seed with a controlled experiment: three EfficientNet-B0 detectors were trained on an identical FaceForensics++ subset (2 000 real + 2 000 fake faces), holding architecture and data fixed and varying only the random seed $\in \{0, 1, 2\}$. PGD perturbations ($\epsilon = 4/255$, $\alpha = 1/255$, 10 steps) were crafted against each variant and cross-evaluated (Figure 5). White-box evasion is 100% on the diagonal; crucially, *cross-seed transfer evasion averages 66.7%* (versus a 23.7% clean baseline) — adversarial perturbations transfer strongly between detectors that differ only in random seed. Combined with the held-out re-evaluation above — where PGD did *not* transfer to a same-architecture detector trained on different data — this isolates the operative factor: **PGD transferability is governed by training-data overlap**

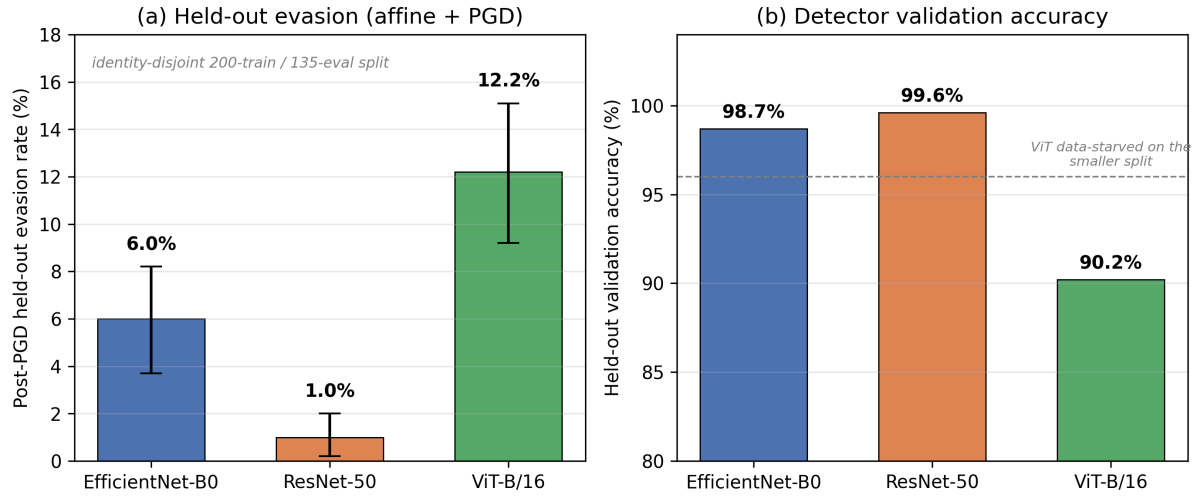


Figure 4: Held-out multi-detector evasion under an identity-disjoint 200-train / 135-eval video split (seed 20260520). (a) Post-PGD evasion on held-out affine-warped fakes with bootstrap 95% CIs — evasion stays low (1–12%) and PGD provides no transfer. (b) Held-out validation accuracy; ViT-B/16 plateaus below the CNNs, data-starved on the smaller split.

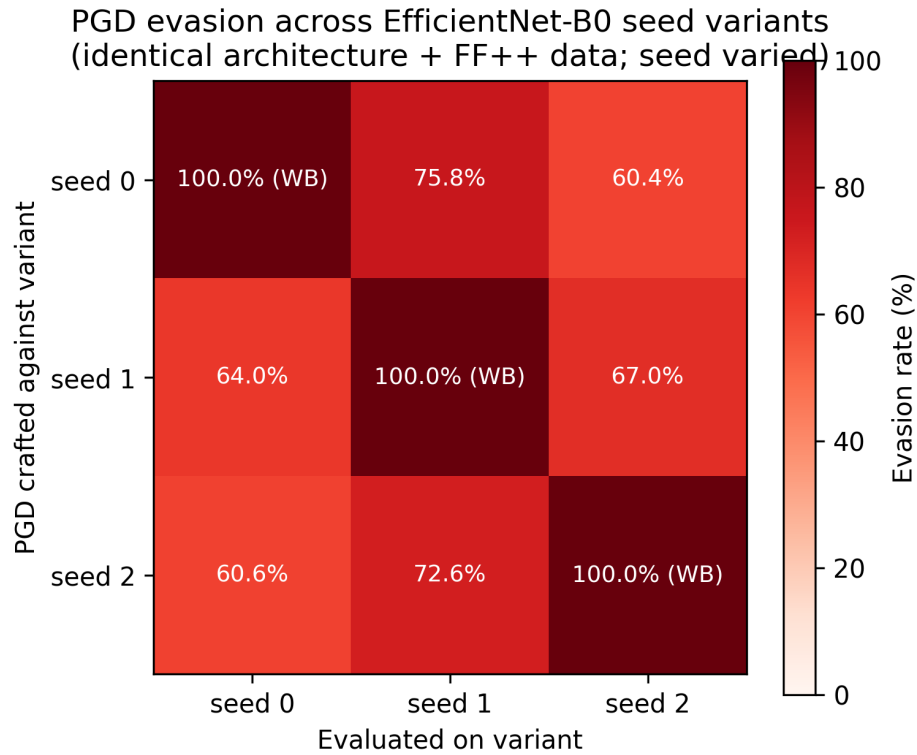


Figure 5: PGD seed-control experiment. Three EfficientNet-B0 detectors trained on an identical FaceForensics++ subset, varying only the random seed. Cell (i, j) is the evasion rate of PGD crafted against variant i and evaluated on variant j . The diagonal (white-box) is 100%; off-diagonal cross-seed transfer averages 66.7% against a 23.7% clean baseline, showing PGD transfers across random seeds when architecture and training data are held fixed.

rather than by random seed. No experiment here varies architecture with training data held fixed, so we do not claim an architecture-independence result. Section 9.5 develops the implication.

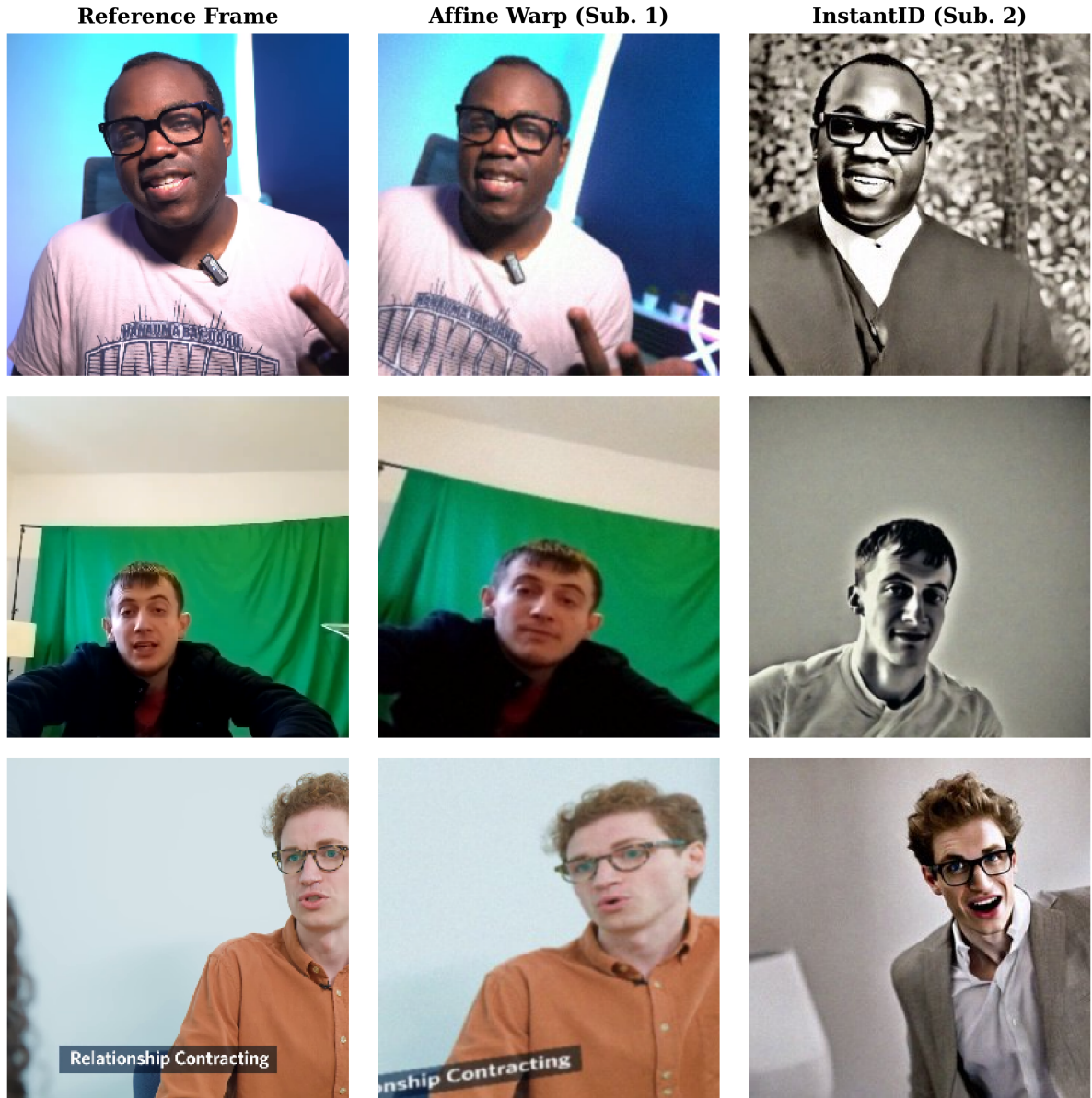


Figure 6: Example generations for three identities. Left: reference frame. Center: affine warp (Submission 1). Right: InstantID (Submission 2). Note InstantID’s tendency toward stylised/desaturated outputs.

Table 6

Frequency-domain metrics (mean over a seeded sample of 200 images per source; seed 20260520).

Source	HF ratio	Noise est.	Laplacian var.
Real (video)	0.678	1.61	799.5
Affine warp	0.605	1.81	136.7
InstantID	0.677	4.43	1050.5

6.4. Frequency Forensics (Generation Side)

To understand *why* detection difficulty varies, we analyse the frequency-domain characteristics of real video frames, affine-warped images, and InstantID outputs (Table 6, Figure 7). This is the Generation-side frequency analysis; a complementary Detection-side analysis is reported in §7.6.

Three observations emerge. First, **real and InstantID HF energy ratios are nearly identical** (0.678 vs. 0.677), with affine warps slightly lower (0.605), confirming that first-order spectral analysis does not

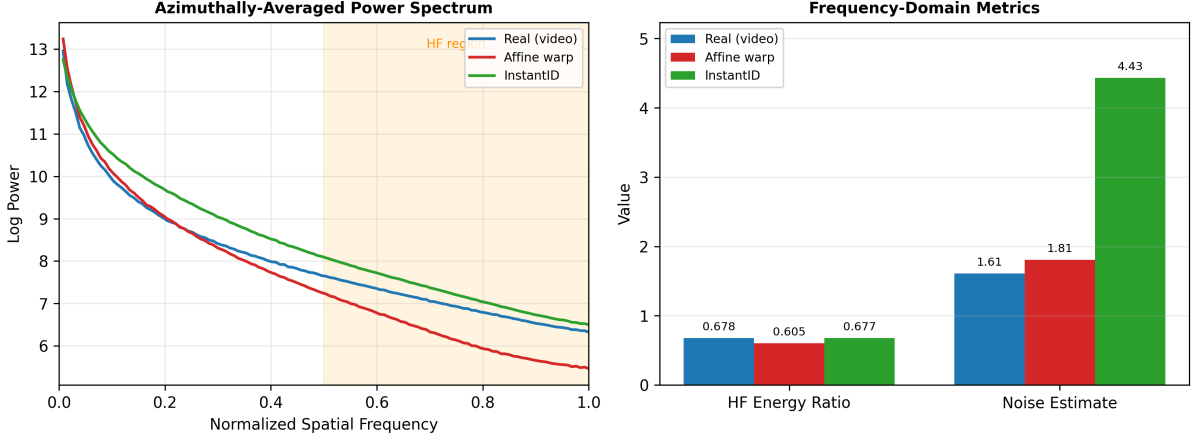


Figure 7: Left: azimuthally-averaged power spectra. All three sources share similar overall spectral shape. Right: real (0.678) and InstantID (0.677) HF energy ratios are nearly identical while affine warps are lower (0.605); noise estimates diverge sharply — InstantID’s 4.43 is $2.7\times$ the real value of 1.61, caused by our $\sigma=4.0$ noise injection.

Table 7

Post-processing noise ablation, single detector-training run. Evasion rate (% classified as real) on 100 raw InstantID images with varying noise σ ; all other post-processing (blur $\sigma=0.3$, JPEG $q=85$) held constant. The five-seed re-run of Section 6.3 reproduces the trend — mean evasion rising from 16% at $\sigma=4.0$ to 30% at $\sigma=0$, with the convolutional detectors noise-sensitive and ViT-B/16 flat — confirming the ablation direction is seed-robust.

σ	EffNet-B0	ResNet-50	ViT-B/16	Mean
0.0	35.0%	50.0%	9.0%	31.3%
0.5	31.0%	51.0%	10.0%	30.7%
1.0	29.0%	44.0%	10.0%	27.7%
1.5	22.0%	38.0%	10.0%	23.3%
2.0	17.0%	31.0%	9.0%	19.0%
4.0	12.0%	3.0%	10.0%	8.3%

cleanly separate diffusion outputs from real frames. Second, our post-processing noise injection ($\sigma=4.0$) is **self-defeating**: it elevates InstantID’s noise estimate to 4.43 — $2.7\times$ the real value of 1.61. Third, InstantID’s **elevated Laplacian variance** (1050 vs. 800 real) indicates sharper edges from the SDXL generation process, providing a secondary detection cue for CNN-based architectures.

6.4.1. Post-Processing Ablation.

To quantify the effect of noise injection, we sweep $\sigma \in \{0, 0.5, 1.0, 1.5, 2.0, 4.0\}$ on 100 InstantID images and evaluate all three detectors (Table 7, Figure 8).

Removing noise entirely ($\sigma=0$) improves mean evasion from 8.3% to 31.3% — a $3.8\times$ improvement. The effect is most dramatic for ResNet-50, whose evasion rate jumps from 3% to 50%: **ResNet-50 was detecting the injected noise, not the deepfake itself**. ViT-B/16 is unaffected ($\sim 9\text{--}10\%$ across all σ values), suggesting it relies on features other than pixel-level noise patterns. Based on this finding, our third submission uses $\sigma=0$ (blur and JPEG compression only, no noise injection), achieving 33.7% (EffNet), 51.6% (ResNet), and 14.1% (ViT) evasion on the full 1 002-image evaluation.

6.5. TTS Model Router Analysis

Table 8 reveals a clear overfitting pattern. Within known speakers, the router achieves 100% of brute-force SECS (0.8177). However, the LOSO evaluation (47.7% for RF, 51.9% for GB) shows that routing accuracy on unseen speakers drops to near the majority baseline.

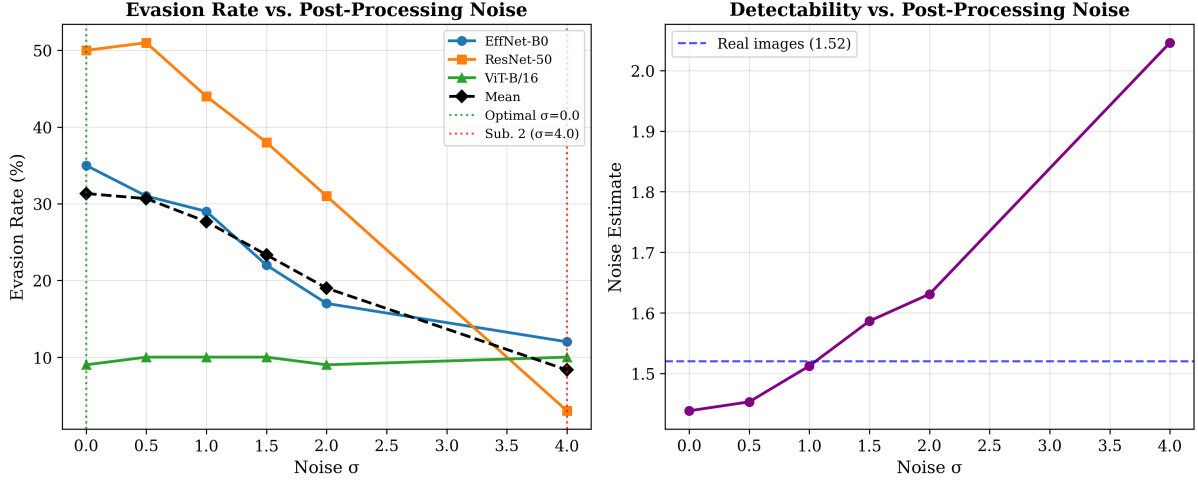


Figure 8: Post-processing noise ablation. Left: evasion rate vs. noise σ . Removing noise ($\sigma=0$) improves mean evasion from 8.3% to 31.3%. ResNet-50 evasion jumps from 3% to 50% — it was detecting the injected noise, not the deepfake. Right: noise estimate converges to the real-image level as $\sigma \rightarrow 0$. The real-image reference here (1.52) is measured on this 100-image ablation sample; Table 6 reports 1.61 on a separate seeded 200-image sample.

Table 8

TTS router evaluation under three protocols.

Protocol	RF Accuracy	GB Accuracy
Majority baseline (always F5-TTS)		47.5%
Stratified 5-fold CV	64.4%	65.4%
Leave-one-speaker-out	47.7%	51.9%
Hold-out 3 speakers	41.1%	44.4%

This demonstrates that TTS model preference is **speaker-specific**: the best model for a given utterance depends more on the speaker-model compatibility than on text-level features (Figures 9 and 10). For practical deployment, this implies per-speaker calibration (running all models on a few calibration utterances per speaker, then using the speaker-level winner for remaining utterances) would outperform a universal router.

Feature Importance. Text features account for 58% of routing importance, audio features 26%, and speaker-aggregate features 16% (Figure 11). The top features are text length (15.4%), reference duration (9.6%), and space density (9.6%). The per-model analysis reveals interpretable patterns: F5-TTS wins on longer utterances (mean 82.6 characters, +14.3% vs. losses) while IndexTTS wins on shorter utterances (71.9 characters, −11.6%) with longer average word length (+4.1%). This suggests F5-TTS’s flow matching architecture handles longer temporal dependencies better, while IndexTTS’s autoregressive approach is more precise on shorter, lexically complex utterances.

6.6. Few-Shot Per-Speaker Calibration

The router’s failure to generalise motivates a simpler and more effective approach (§4.5). With K calibration utterances per speaker we select the best TTS model and use it for all remaining utterances. This is a straightforward baseline that any practitioner could implement; we evaluate it rigorously to quantify its effectiveness.

Table 9 shows that even $K=1$ (a single calibration utterance) outperforms always-F5-TTS, and $K=3$ achieves 98.7% of brute-force SECS while reducing compute by 60% (36 vs. 90 inferences per speaker). The 95% CI for $K=3$ is [0.803, 0.811], entirely above the always-F5-TTS baseline of 0.799 (Figure 12).

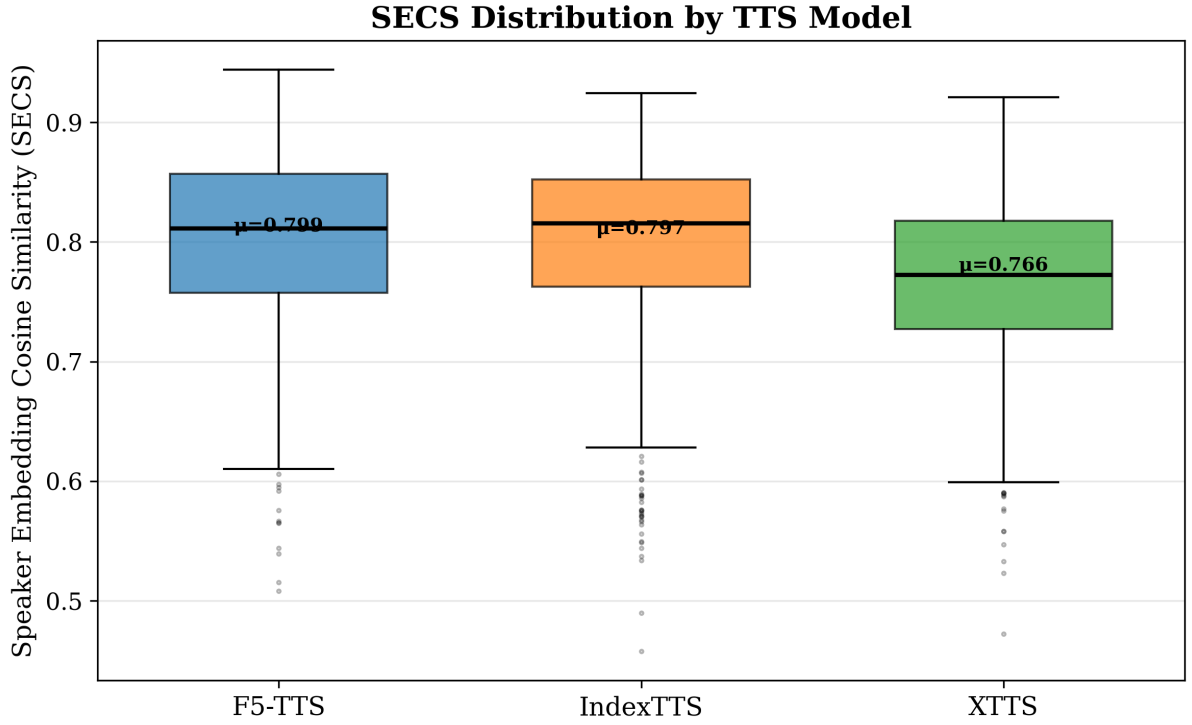


Figure 9: SECS distribution by TTS model. F5-TTS and IndexTTS achieve similar mean SECS ($\mu=0.799$ and 0.797), while XTTS lags ($\mu=0.766$). The overlapping distributions explain the near-even win split (47.5% vs. 44.0%).

Table 9

Few-shot per-speaker calibration (100 Monte Carlo trials, bootstrap 95% CIs). SECS measured on held-out utterances only. “% of BF” = percentage of brute-force ensemble SECS achieved.

Method	SECS	95% CI	% of BF	Compute
Brute force	0.8177	—	100.0%	3×all
$K=3$ calibration	0.8068	[.803, .811]	98.7%	36/spk
$K=5$ calibration	0.8082	[.804, .811]	98.8%	40/spk
$K=10$ calibration	0.8096	[.806, .813]	99.0%	50/spk
Always F5-TTS	0.7989	—	97.7%	1×all
Random	0.7849	—	96.0%	1×all

The marginal gain diminishes beyond $K=5$, suggesting 3–5 calibration utterances is the practical sweet spot.

6.7. ControlNet-478 vs. InstantID

Table 10 compares the ControlNet-478 trained on full 478-point landmarks against InstantID’s 5-keypoint approach on 5 test identities.

ControlNet-478 generates recognisable faces but achieves **18× worse landmark accuracy than InstantID** (0.167 vs. 0.009). Three factors account for the gap: (1) InstantID was pretrained on millions of face images while CN-478 trains on only 1 601 pairs for 5 000 steps; (2) InstantID injects a 512-dimensional identity embedding via IP-Adapter, providing a strong identity prior that CN-478 lacks entirely; and (3) the from-scratch UNet initialisation has not fully adapted to dense 478-landmark conditioning.

The training itself achieved zero NaN losses across all 5 000 steps with final loss 0.0885 (Figure 13), confirming that the architecture is stable and that the full 478-point MediaPipe Face Mesh can be used as a viable conditioning signal for ControlNet. To our knowledge, this is the first ControlNet trained on

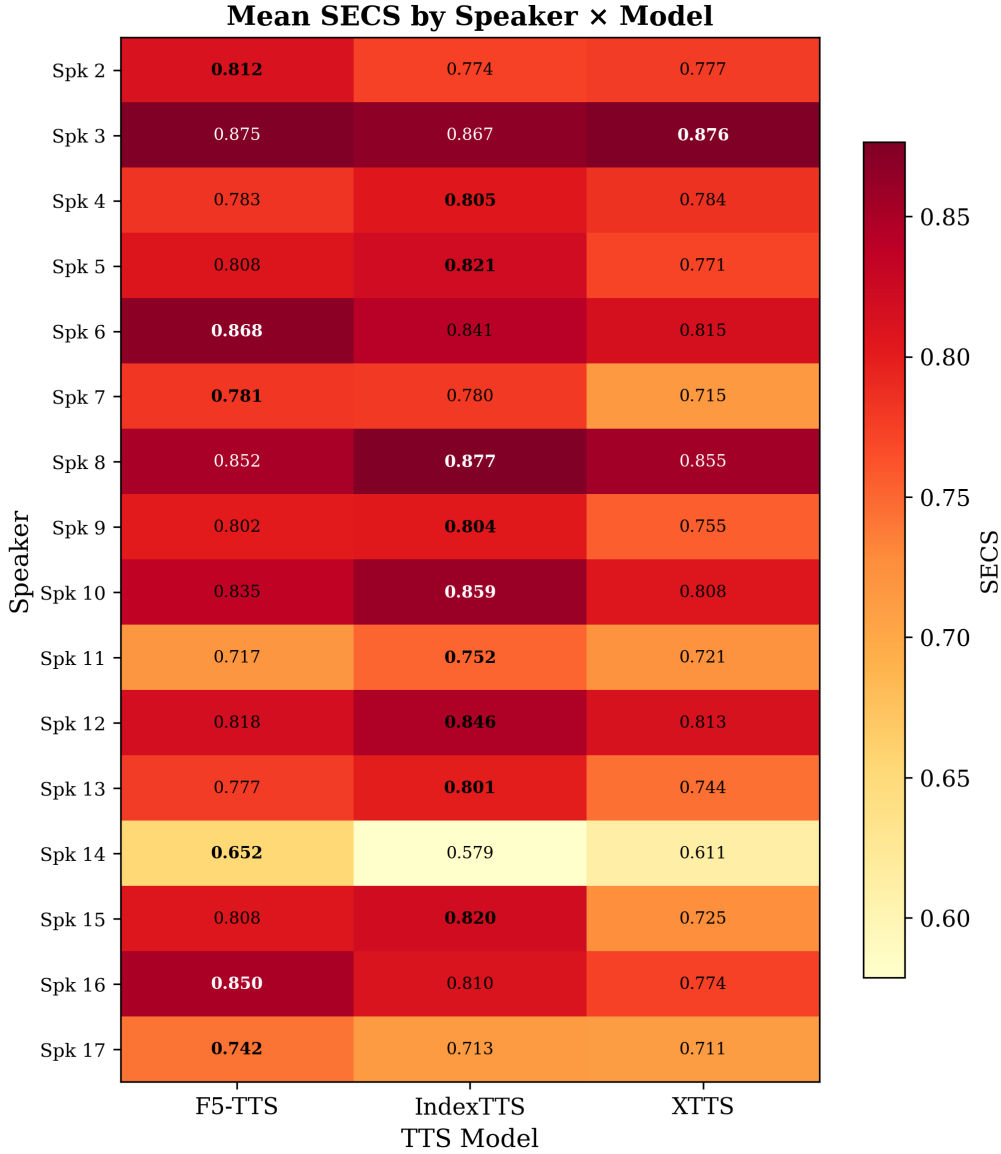


Figure 10: Mean SECS by speaker $\times$ model. Bold values indicate the best model per speaker. The winning model varies across speakers (F5-TTS wins 6, IndexTTS wins 9, XTTS wins 1), confirming that model preference is speaker-specific rather than universal.

the complete 478-point landmark set.

7. Detection Results

7.1. Per-Detector Calibration on the Gen-Task Set

We evaluate each of the eight candidate image detectors on the 394-sample calibration set using six diagnostic metrics: AUC with bootstrap 95% confidence interval (1 000 resamples), Expected Calibration Error (ECE, 15-bin Guo et al. [14]), Brier score, F1 at $\tau = 0.5$ for the fake class, Equal Error Rate (EER), and Matthews correlation coefficient (MCC). Table 11 reports per-detector results.

Drop decisions are unambiguously supported. ash12321 has AUC 0.574 with lower-CI 0.510 (crossing the 0.5 chance line), and freqms_b4 has AUC 0.523 with F1 of 0.000 at $\tau = 0.5$ (it predicts the same class for every sample on this distribution). Both are dropped from v2.

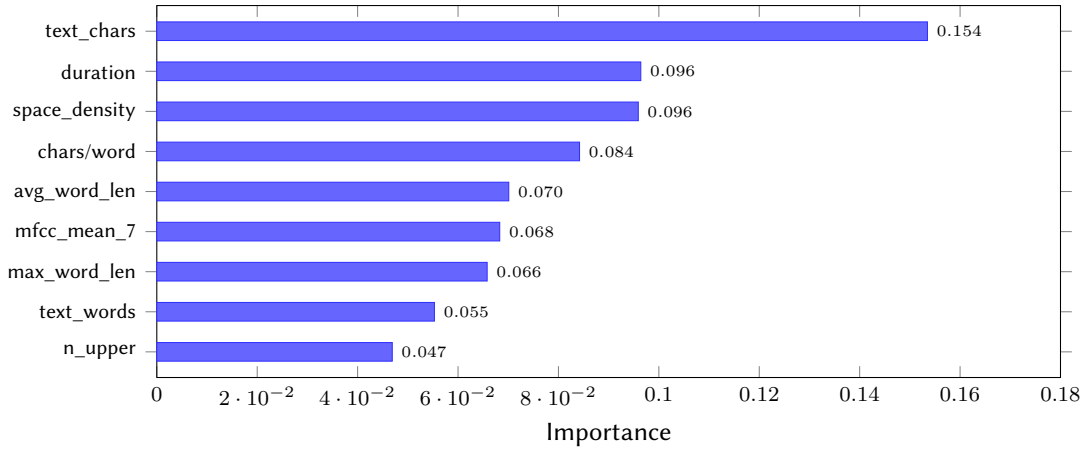


Figure 11: Top 9 features for TTS model routing by importance. Text features (blue) dominate, with `text_chars` being the strongest predictor. `duration` (the reference-audio clip duration) is the only audio feature in the top 5.

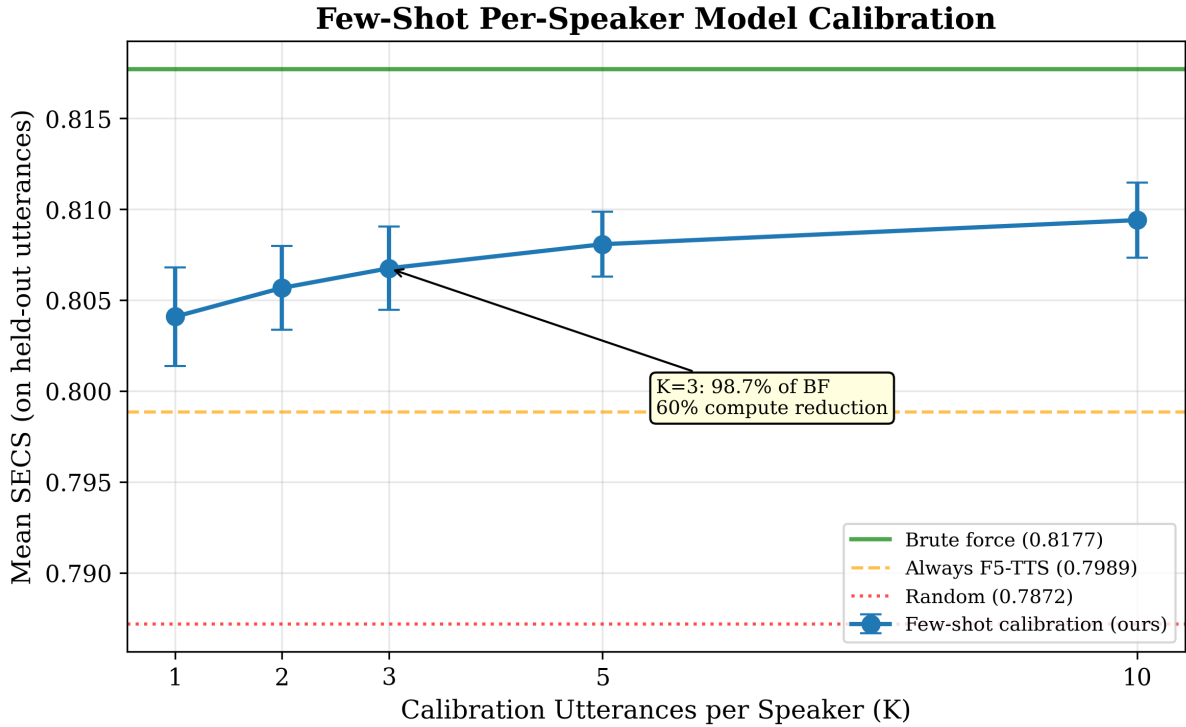


Figure 12: Few-shot per-speaker calibration: SECS vs. calibration utterances K , with 95% CIs from 100 Monte Carlo trials. $K=3$ achieves 98.7% of brute-force SECS at 60% compute reduction.

Pairwise DeLong significance. We compute the DeLong–DeLong–Clarke–Pearson pairwise AUC test across all 28 detector pairs, with Bonferroni correction at $\alpha = 0.05/28 = 0.00179$. **23 of 28 pairs are significant** after correction. The 5 non-significant pairs cluster naturally by performance tier: `umm_maybe` $\leftrightarrow$ `dima806` (both AUC ≈ 0.94 , top tier), `prithivMLmods_v2` $\leftrightarrow$ `Organika` (both AUC ≈ 0.87 , second tier), and `ash12321` $\leftrightarrow$ `fraqms_b4` (both at chance, dropped tier). This three-tier structure mirrors the DeLong test’s natural cluster boundaries.

Heterogeneous miscalibration is a finding in itself. Figure 14 shows the six reliability diagrams for the v2-kept detectors. Six distinct miscalibration patterns emerge:

1. `prithivMLmods_v2`: flat-near-zero, then steep S-curve crossing the diagonal at ~ 0.8 . Under-

Table 10

Landmark distance comparison: ControlNet-478 (5 000 steps, 1 601 training pairs) vs. InstantID (pretrained on millions of images with identity embeddings). Lower is better.

Image	CN-478	InstantID
ID_0_GT_frame1	0.0775	0.0077
ID_0_GT_frame2	0.2663	0.0065
ID_0_GT_frame3	0.1806	0.0204
ID_100_GT_frame1	0.2036	0.0045
ID_100_GT_frame2	0.1048	0.0060
Mean	0.1666	0.0090

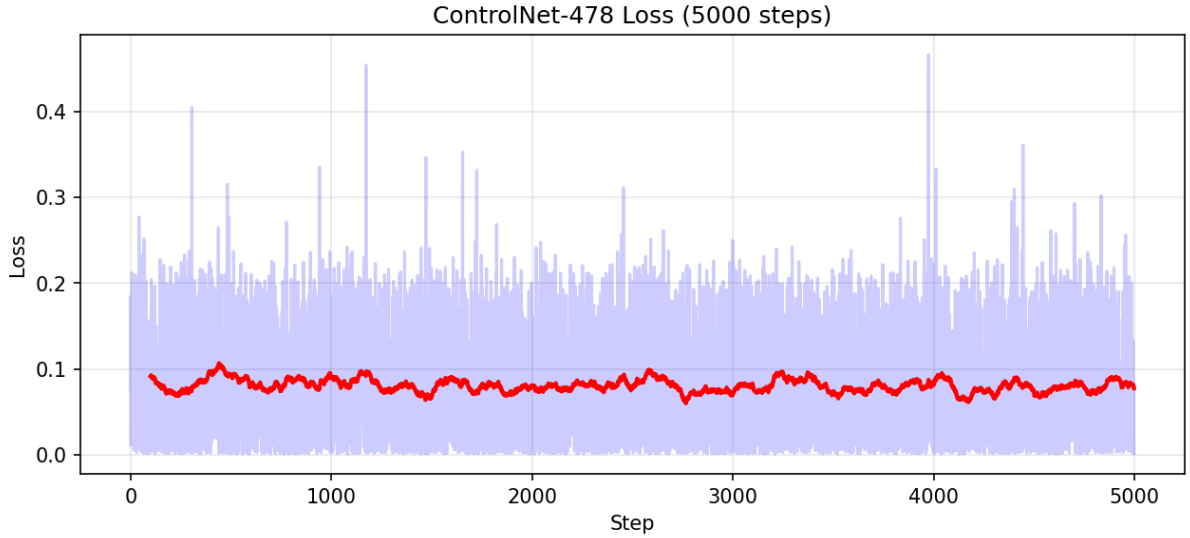


Figure 13: ControlNet-478 training loss over 5 000 steps. The smoothed curve (red) converges to 0.0885 with zero NaN losses, confirming architectural stability of 478-point landmark conditioning.

Table 11

Per-detector calibration statistics on the 394-sample calibration set (194 real + 200 user-fake from our Generation outputs). AUC reported with bootstrap 95% CI (1 000 resamples). Verdict column indicates v2 inclusion decision; detectors marked HURTS were dropped.

Detector	AUC [95% CI]	ECE	BS	F1	EER	Verdict
prithivMLmods_v2	0.870 [0.832, 0.907]	0.259	0.159	0.815	0.208	HELPS
prithivMLmods_siglip	0.735 [0.683, 0.784]	0.305	0.232	0.673	0.337	HELPS
dima806	0.944 [0.921, 0.967]	0.460	0.151	0.851	0.129	HELPS
umm_maybe	0.948 [0.927, 0.966]	0.323	0.116	0.853	0.137	HELPS
Organika	0.866 [0.828, 0.903]	0.434	0.171	0.799	0.208	HELPS
ash12321	0.574 [0.510, 0.634]	0.311	0.379	0.586	0.408	HURTS
proxy_eff_b0	0.621 [0.563, 0.679]	0.417	0.381	0.439	0.410	NEUTRAL
freqms_b4	0.523 [0.477, 0.571]	0.493	0.508	0.000	0.468	HURTS

confident at low end, well-calibrated at high end.

2. prithivMLmods_siglip: consistently above diagonal. Overconfident in low predictions (says $P(\text{fake}) = 0.2$ when truth is 0.4).
3. dima806: bimodal (flat at 0, sharp jump to 1). Extreme polarisation with almost no middle-confidence predictions.
4. umm_maybe: below diagonal throughout. Underconfident / conservative, predicting $P(\text{fake}) = 0.5$ when truth is 0.25.

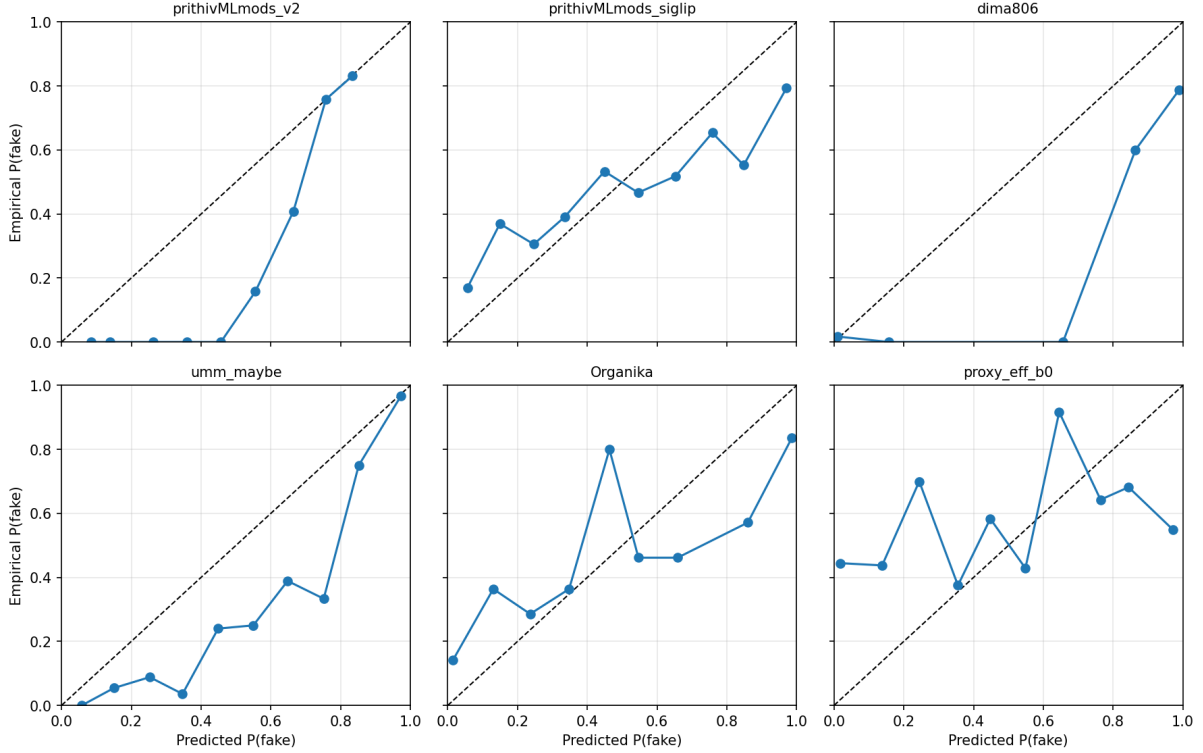


Figure 14: Reliability diagrams for the six detectors retained in v2. Each panel shows predicted $P(\text{fake})$ vs. empirical $P(\text{fake})$ in 10 score-bins (minimum 5 samples per bin). All six detectors are miscalibrated, with six qualitatively distinct patterns. `proxy_eff_b0` (bottom-right) is non-monotonic, justifying its cross-task-only inclusion rather than performance-based inclusion.

5. `Organika`: severe under-confidence with mid-range spike at $0.5 \rightarrow 0.8$. Detector “saw evidence but doesn’t trust itself.”
6. `proxy_eff_b0`: **non-monotonic chaos**; oscillates from 0.45 to 0.92 to 0.55 across the score range. Not just miscalibrated; fundamentally not a reliable probability.

Raw-score ECE values range from 0.26 to 0.49 across the six kept detectors — roughly an order of magnitude above the ECE values ($\lesssim 0.05$) commonly used as a rule-of-thumb for a well-calibrated classifier. This heterogeneous miscalibration explains mechanistically why v1’s rank-norm ensemble works poorly: errors do not cancel because they are asymmetric in different directions for different score ranges. The v2 LR meta-learner can learn *per-detector* corrective transforms simultaneously, which is the mechanism behind its improvement over v1.

7.2. Orientation Analysis

7.2.1. Within-Distribution: Labelled, Anchor, Bootstrap, Subgroup

Labelled orientation on Gen-task. For each detector d , we compute $\text{AUC}(d) = \text{roc_auc_score}(y, s_d)$ on the calibration set. All eight detectors achieve $\text{AUC} > 0.5$ on Gen-task, meaning *zero of eight* detectors require orientation flipping on our distribution.

Anchor procedure agreement. The unsupervised anchor procedure (Section 5.4) agrees with the labelled orientation on 6 of 8 detectors. The two disagreements are `ash12321` (anchor says inverted, labelled says normal; median $\rho = -0.054$) and `freqms_b4` (anchor says inverted, labelled says normal; median $\rho = -0.001$). Both detectors have near-zero inter-detector correlations, making the anchor median unstable. Notably, these are also the two HURTS detectors dropped from v2. *The anchor procedure fails on exactly the detectors that should not be in the ensemble.*

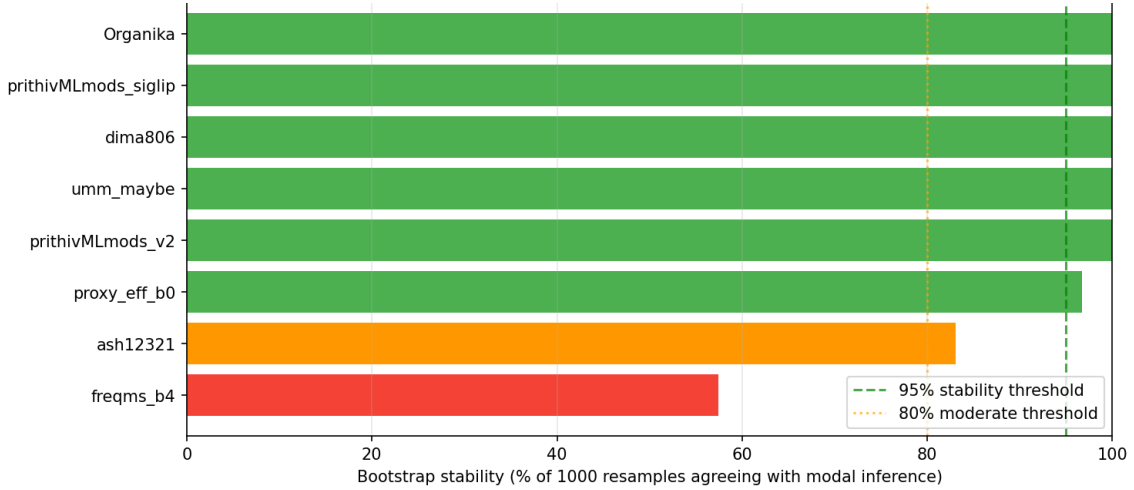


Figure 15: Per-detector bootstrap stability of the anchor procedure. Green bars indicate $\geq 95\%$ stability; orange $\geq 80\%$; red $< 80\%$. Five of six v2-kept detectors are 100% stable; proxy_eff_b0 at 96.7%. The two HURTS detectors (ash12321, freqms_b4) are unstable, mirroring their poor standalone calibration performance.

Bootstrap stability. We resample the calibration set 1 000 times (with replacement) and re-infer orientation on each resample. Per-detector stability is the percentage of resamples agreeing with the modal inference (Figure 15). Five of the six v2-kept detectors achieve 100% stability; proxy_eff_b0 is 96.7% stable; the two HURTS detectors are unstable (ash12321 83.1%, freqms_b4 57.4%). The procedure’s reliability tracks detector reliability.

Subgroup-split robustness. We split the 200 fake samples into two subgroups by index (subgroup A = fake_0000–0099, subgroup B = fake_0100–0199) corresponding to our InstantID and affine-warp generator outputs respectively. Anchor inference agrees across the two subgroups for 7 of 8 detectors; only ash12321 (the marginal HURTS detector) flips between subgroups. The detector pairwise-correlation structure underlying the anchor procedure is shown in Figure 16.

Threshold sensitivity. The default anchor-procedure threshold ($\rho \geq 0$ implies normal) yields 6/8 agreement. Lowering to $\rho \geq -0.10$ produces 8/8 agreement on Gen-task by tolerating the near-zero correlations of the two unstable detectors. This suggests a methodological refinement: anchor procedures should use a small negative tolerance rather than strict zero, to handle detectors whose median correlation is genuinely undefined.

7.2.2. Cross-Distribution: UADFV vs Gen-Task

A suggestive but small-sample observation from our orientation analysis is **apparent cross-distribution variability**, which we report with the caveats developed below. We cross-reference per-detector AUCs from our preliminary V2-1 analysis on the 1 000-sample UADFV control with the labelled Gen-task AUCs from Table 11, recovering raw (pre-flip) AUC values for each distribution. Table 12 reports the cross-distribution table.

Four detectors (dima806, umm_maybe, ash12321, freqms_b4) have raw AUC < 0.5 on UADFV but raw AUC ≥ 0.5 on Gen-task. Only two of these are clean flips: dima806 (0.460 $\rightarrow$ 0.944) and umm_maybe (0.394 $\rightarrow$ 0.948), both with large AUC swings and 100% bootstrap-stability. The other two are not. ash12321 sits at raw AUC 0.499 on UADFV — exactly chance — and freqms_b4 is within 0.03 of chance on *both* distributions (0.469 / 0.523), only 57.4% bootstrap-stable, with an effectively uninitialised head (§5.2). A detector at AUC ≈ 0.5 has no determinate orientation, so labelling it “flipped” reflects thresholding noise around chance rather than a genuine reversal. Read honestly, the evidence is **two clean flips among the six usable detectors**, not four of eight. This is *consistent with* public detector orientations being distribution-dependent and motivates unsupervised distribution-aware inference such as the anchor procedure, but two caveats bound it: the comparison is a single distribution

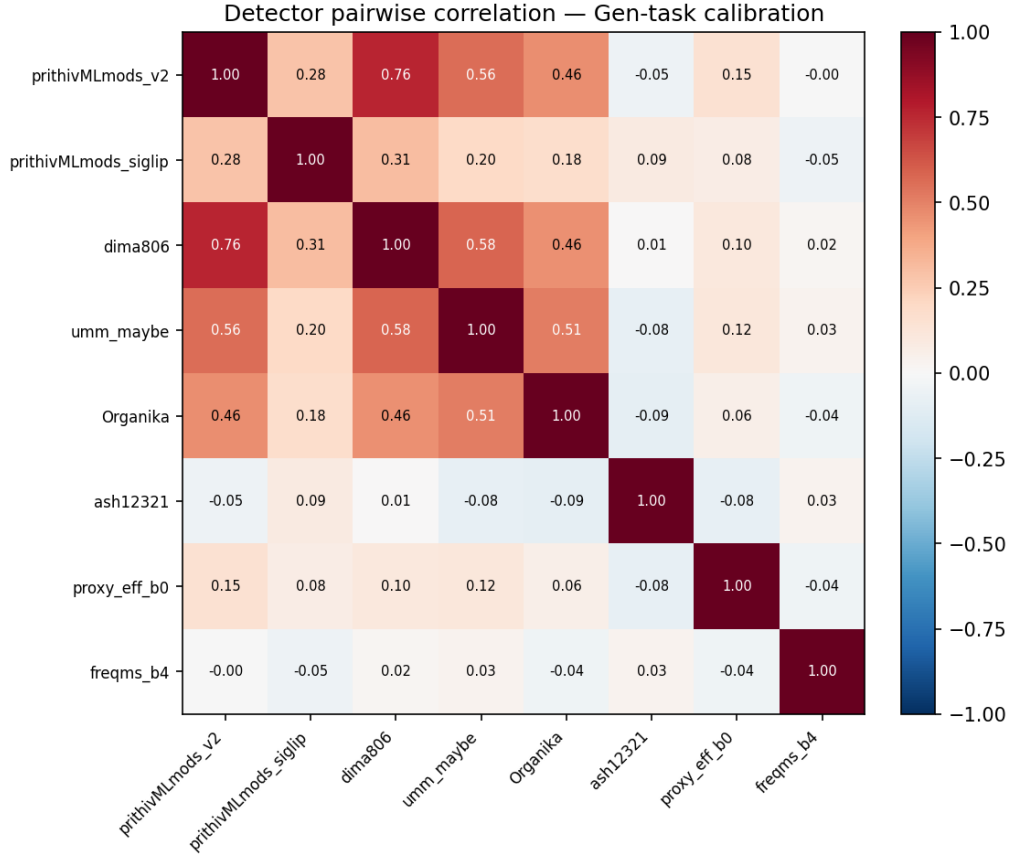


Figure 16: Detector pairwise correlation heatmap on the Gen-task calibration set. Three natural tiers are visible: a strongly correlated cluster of five mainstream detectors ($\rho = 0.18\text{--}0.76$), proxy_eff_b0 weakly correlated with everyone ($\rho \approx 0.06\text{--}0.15$), and the two HURTS detectors ash12321 and freqms_b4 essentially decoupled ($\rho \approx 0$ with everything).

Table 12

Cross-distribution orientation evidence. Eight detectors evaluated on UADFV ($n = 1\,000$) and our Gen-task calibration set ($n = 394$). Raw AUC reported pre-flip. **Four of eight detectors show opposite raw-AUC orientations on the two distributions** — two cleanly, two at near-chance AUC (see text) — consistent with no fixed orientation table reliably generalising across distributions. “Stable%” is the Gen-task bootstrap-stability percentage from Figure 15.

Detector	UADFV ($n = 1\,000$)		Gen-task ($n = 394$)		Stable%
	Raw AUC	Flip?	Raw AUC	Flip?	
prithivMLmods_v2	0.734	No	0.870	No	100.0
prithivMLmods_siglip	0.600	No	0.735	No	100.0
dima806 †	0.460	Yes	0.944	No	100.0
umm_maybe †	0.394	Yes	0.948	No	100.0
Organika	0.504	No	0.866	No	100.0
ash12321 †	0.499	Yes	0.574	No	83.1
proxy_eff_b0	0.685	No	0.621	No	96.7
freqms_b4 †	0.469	Yes	0.523	No	57.4

† Detector shows opposite raw-AUC orientation between UADFV and Gen-task; ash12321 and freqms_b4 sit at near-chance AUC and have no determinate orientation (see text).

pair, and UADFV is a 2019-era GAN/landmark-manipulation dataset while our Gen-task fakes are diffusion-era, so distribution and generator-era covary and the flips cannot be attributed to distribution shift alone. We therefore treat this as a small-sample, single-pair observation rather than a definitive

Table 13

v1 vs v2 prediction-level statistical comparison (Phase 1: no leaderboard labels). Pearson r reported with 95% CI via Fisher’s z -transformation. McNemar test reports symmetry of $v1 \rightarrow v2$ prediction flips at $\tau = 0.5$; large $b - c$ asymmetry indicates v2 systematically shifted predictions in one direction.

Statistic	Images ($n = 24\,404$)	Audio ($n = 11\,520$)
Pearson r [95% CI]	0.087 [0.075, 0.100]	0.366 [0.350, 0.381]
KS statistic (p)	0.516 ($p \approx 0$)	0.521 ($p \approx 0$)
Agreement @ $\tau = 0.5$	52.5%	68.9%
McNemar b ($v1$ =fake, $v2$ =real)	6 202	1 988
McNemar c ($v1$ =real, $v2$ =fake)	5 382	1 600
McNemar χ^2 (p)	57.90 ($p = 2.8 \times 10^{-14}$)	41.74 ($p = 1.0 \times 10^{-10}$)
v1 positive rate	51.6%	47.9%
v2 positive rate	48.2%	44.5%

finding; a within-era control is needed to isolate the effect.

7.2.3. A Cautionary Methodology Lesson

We report a self-acknowledged failed orientation correction attempt from our earlier V2-3-fix analysis. That attempt used three detectors (`prithivMLmods_v2`, `prithivMLmods_siglip`, `proxy_eff_b0`) as anchors and inferred that `dima806`, `Organika`, and (incorrectly) `umm_maybe` required flipping on Gen-task. The procedure’s self-assessment was HURTS ($v1 \leftrightarrow v2$ correlation barely moved from -0.087 to $+0.085$).

The cause, made visible by the reliability diagrams in Figure 14, is that `proxy_eff_b0` is *non-monotonic*: correlations against it do not carry orientation information. Using a non-monotonic detector as an anchor pollutes the median correlation for all other detectors. **The lesson is that anchor procedures should screen anchor candidates for monotonicity via reliability-diagram inspection before use.** Our v2 anchor set excludes `proxy_eff_b0` and uses only monotonic detectors as anchors.

7.3. v1 vs v2 Prediction-Level Comparison

We compare v1 and v2 prediction surfaces on the held-out test set (without ground-truth labels, since official scores are pending):

Low image Pearson is structural, not anomalous. v1 outputs are essentially binary (mass at $\{0, 1\}$, Figure 17 left column), while v2 outputs are continuous probabilities. The low correlation reflects both architectural change (rank-norm to LR meta-learner) and output-type difference (binary to continuous). Future submissions could enforce continuous output from v1 by removing the rank-norm binarisation, which would yield a more interpretable comparison.

Audio Pearson is meaningfully higher (+0.366) than image (+0.087). Audio v2 is a simple-average ensemble of the same three detectors used in v1 (no meta-learner), so the cross-version drift is dominated by orientation-correction adjustments rather than architectural change. This is the expected pattern.

McNemar asymmetry is consistent across modalities. Both $b - c$ values indicate v2 shifted approximately 3.4% of predictions from “fake” to “real” net of reverse flips, on both modalities. This is consistent with v2 being better-calibrated to the $\sim 50/50$ base rate than v1’s binary outputs.

7.4. Ablation Studies

We perform leave-one-detector-out ablation on the six v2-kept image detectors, with 5-fold cross-validation, and compare four meta-learner architectures plus the best single detector.

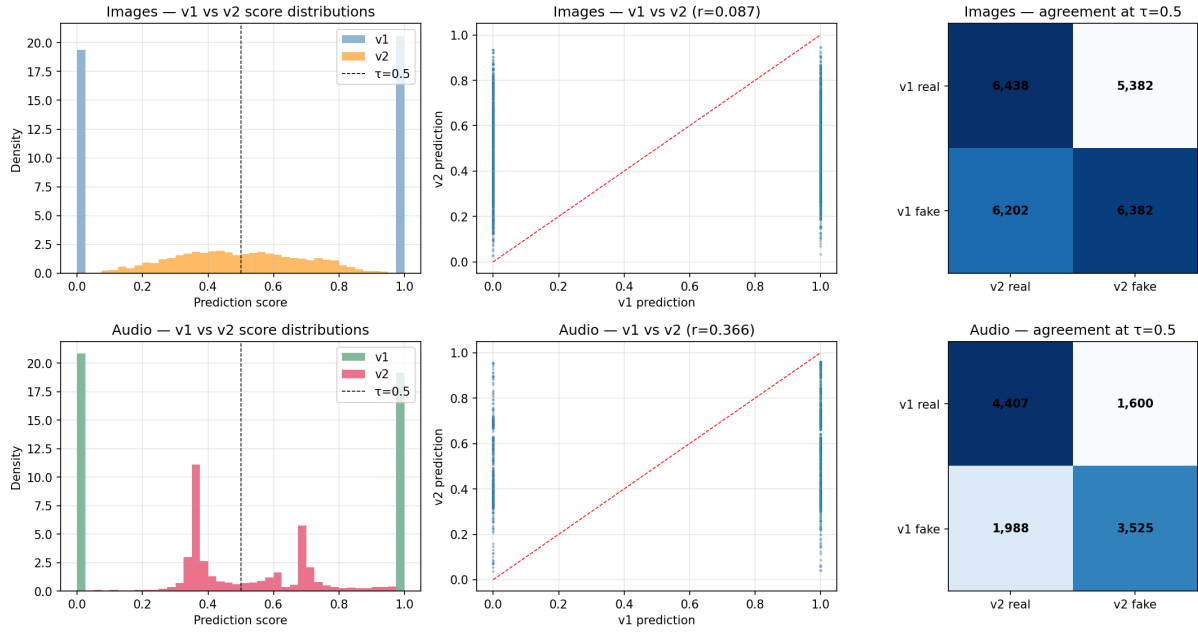


Figure 17: v1 vs v2 prediction-level comparison. *Left:* histograms of score distributions. v1 (blue/green) is bimodal at $\{0, 1\}$; v2 (orange/red) spans the full $[0, 1]$ range, with audio v2 showing characteristic three-detector averaging structure (peaks at $\sim 0.35, 0.55, 0.65$). *Middle:* scatter plots demonstrate low image correlation ($r = 0.087$) and moderate audio correlation ($r = 0.366$). *Right:* agreement quadrants at $\tau = 0.5$.

Table 14

Leave-one-out ablation of the six v2-kept image detectors. Each row reports the 5-fold CV AUC obtained by removing the indicated detector from the LR meta-learner. ΔAUC is the drop relative to the full 6-detector baseline AUC of 0.9896.

Detector removed	AUC without	ΔAUC	Verdict
umm_maybe	0.9710	0.0186	HELPFUL
dima806	0.9845	0.0051	HELPFUL
Organika	0.9858	0.0038	REDUNDANT
proxy_eff_b0	0.9886	0.0010	REDUNDANT
prithivMLmods_siglip	0.9887	0.0009	REDUNDANT
prithivMLmods_v2	0.9888	0.0008	REDUNDANT

7.4.1. Per-Detector Contribution (LOO)

Ensemble is robust to individual detector loss. Only umm_maybe (AUC 0.0186) and dima806 (AUC 0.0051) produce non-trivial drops when removed. The remaining four detectors classify as REDUNDANT ($\text{AUC} \leq 0.004$); the leave-one-out ranking is reported in Table 14. This is not a weakness; it demonstrates that v2 performance does not depend on any single detector being uniquely strong, and that the ensemble degrades gracefully under detector loss.

The proxy_eff_b0 redundancy confirms the cross-task framing. The user-trained proxy detector contributes $\text{AUC} = 0.0010$ (*effectively zero* performance contribution). Its inclusion in v2 is therefore explicit cross-task threat-model consistency rather than performance-driven: it is the same detector the InstantID PGD attack was optimised against in Section 4.4, and its presence in the Detection ensemble preserves a coherent attack-defence pairing across the two subtasks.

Table 15

Meta-learner architecture comparison on the 6-detector feature set, 5-fold cross-validated AUC. The LR meta-learner in v2 marginally outperforms XGBoost and substantially outperforms rank-normalised averaging (v1’s architecture) and simple averaging. The LR meta-learner’s AUC has a bootstrap 95% CI of [0.9833, 0.9963] (1 000 resamples).

Method	AUC	std (5-fold)
Logistic regression (v2)	0.9896	0.0066
XGBoost	0.9878	0.0065
Rank-normalised average (v1)	0.9850	—
Simple average	0.9844	—
Best single (umm_maybe)	0.9481	—

7.4.2. Meta-Learner Architecture Comparison

Calibration matters more than meta-learner architecture. LR vs. XGBoost differ by 0.0018 AUC (within std). v1 rank-norm to v2 LR meta-learner produces 0.0046 AUC improvement (*2.6 times the LR-vs-XGBoost gap*), indicating that the calibration step is the dominant intervention, not the meta-learner choice (Table 15). To quantify the LR meta-learner’s uncertainty directly, we bootstrap its AUC over 1 000 resamples of the calibration set, obtaining a 95% CI of [0.9833, 0.9963]; the rank-norm (v1) AUC of 0.9850 lies at the lower edge of this interval, so the v1→v2 gain, while consistent in direction across all resamples, is modest in magnitude.

7.4.3. LLM Tier Contribution

The v2 image pipeline forwards the top- $K = 500$ most uncertain meta-learner predictions to a multi-modal LLM (Claude Sonnet 4.6) and blends the verdict back (§5.6). We characterise this tier’s contribution with two experiments on the calibration set.

Paired bootstrap: the LLM tier is statistically neutral. We compare the meta-learner-only AUC against the LLM-augmented AUC under a paired bootstrap (1 000 resamples, identical resample indices for both systems so the comparison is paired). The mean difference is $\Delta\text{AUC} = 0.000$ with a 95% CI of $[-0.005, +0.005]$, and the LLM-augmented system wins on only 51.6% of resamples — indistinguishable from a coin flip. On our calibration set, the LLM tier neither helps nor hurts the ensemble AUC at a level we can measure.

K -sweep: no operating point recovers a gain. To rule out that $K = 500$ was simply a poor choice, we sweep $K \in \{0, 25, 50, 98, 150, 200, 300, 394\}$ and recompute the blended AUC at each (Figure 18). For this post-hoc analysis the uncertain samples were re-scored with Claude Opus 4.7 — a more capable model than the Sonnet 4.6 used in the original submission — so the sweep additionally tests whether a stronger LLM would have helped. The best point, $K = 50$, yields +0.0017 AUC, but its confidence interval overlaps the $K = 0$ (no-LLM) baseline; beyond $K \approx 150$ the blended AUC declines monotonically as the LLM tier overrides correct meta-learner predictions, reaching an LLM-only AUC of 0.890. No operating point produces a statistically reliable improvement.

Interpretation. We retain the description of the LLM tier as part of the submitted v2 system for completeness and reproducibility, but we do not claim it as a source of measurable gain. Its effect is neutral on our calibration distribution; whether a larger or differently-prompted LLM tier would help on the test distribution is unresolved (§10). This is a deliberately conservative reading: a tier we shipped turns out, on controlled evaluation, not to earn its place on performance grounds.

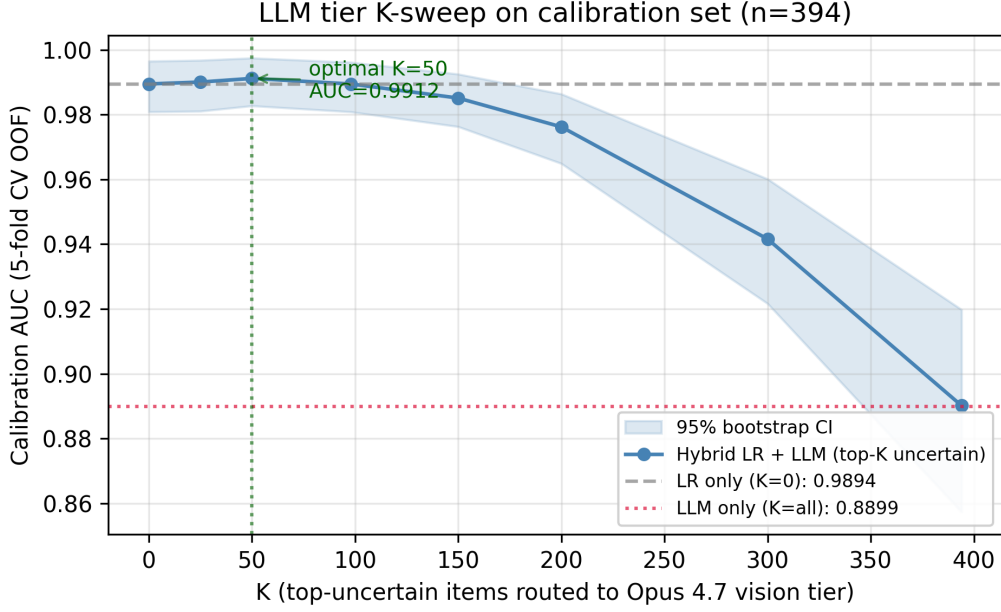


Figure 18: LLM-tier K -sweep. Blended meta-learner + LLM AUC as a function of the number K of most-uncertain samples re-scored by the LLM. The dashed line is the $K = 0$ (meta-learner only) baseline. The best point ($K = 50$) is within the baseline’s confidence interval; AUC declines for $K \gtrsim 150$. The LLM tier does not yield a reliable improvement at any K .

Table 16

Calibration-size sensitivity of the LR meta-learner. Mean AUC and standard deviation across 30 random subsets of indicated size. The LR meta-learner is calibration-size insensitive, achieving $\text{AUC} \approx 0.988$ from $n = 50$ to $n = 300$ calibration samples.

$n_{\text{calibration}}$	Mean AUC	std AUC
50	0.9885	0.0024
100	0.9886	0.0026
200	0.9902	0.0038
300	0.9879	0.0070

7.4.4. Calibration-Size Sensitivity

50 calibration samples suffice. The standard deviation at $n = 300$ is higher because 5-fold CV with 300 samples leaves only ~ 60 evaluation samples per fold; performance is essentially identical to $n = 50$. This eliminates concerns that our 394-sample calibration set is a bottleneck, and informs deployment: a new calibration set of 50 user-generated fakes plus 50 reals is sufficient to train an effective meta-learner (Table 16).

7.5. Threshold Sensitivity

We sweep the operating threshold $\tau \in [0.30, 0.70]$ on the simple-average orientation-corrected ensemble.

Optimal $\tau = 0.60$ yields +2.5 F1 points over submitted $\tau = 0.50$. F1 at the submitted threshold is 0.908; at the calibration-optimal threshold $\tau = 0.60$ it is 0.933 (Table 17). This was chosen pre-submission without leveraging calibration-set tuning. Reported here as a methodology refinement for future submissions: threshold tuning is essentially free post-calibration but produces a non-trivial F1 improvement. ROC and precision-recall curves for the individual detectors and the ensemble are shown in Figure 19.

Table 17

Threshold sensitivity of the v2 simple-average ensemble on the calibration set after orientation correction. F1 reported for the fake class. Submitted v2 uses $\tau = 0.50$; F1 is maximised at $\tau = 0.60$ (*). A pre-submission threshold tuning would have yielded an estimated +2.5 F1 points.

τ	F1 (fake)	Accuracy	FPR	FNR
0.30	0.7984	0.7443	0.5179	0.0000
0.35	0.8386	0.8051	0.3949	0.0000
0.40	0.8547	0.8278	0.3487	0.0000
0.45	0.8844	0.8684	0.2615	0.0050
0.50	0.9078	0.8987	0.1897	0.0150
0.55	0.9268	0.9241	0.1026	0.0500
0.60 *	0.9333	0.9342	0.0410	0.0900
0.65	0.8822	0.8911	0.0205	0.1950
0.70	0.8152	0.8405	0.0103	0.3050

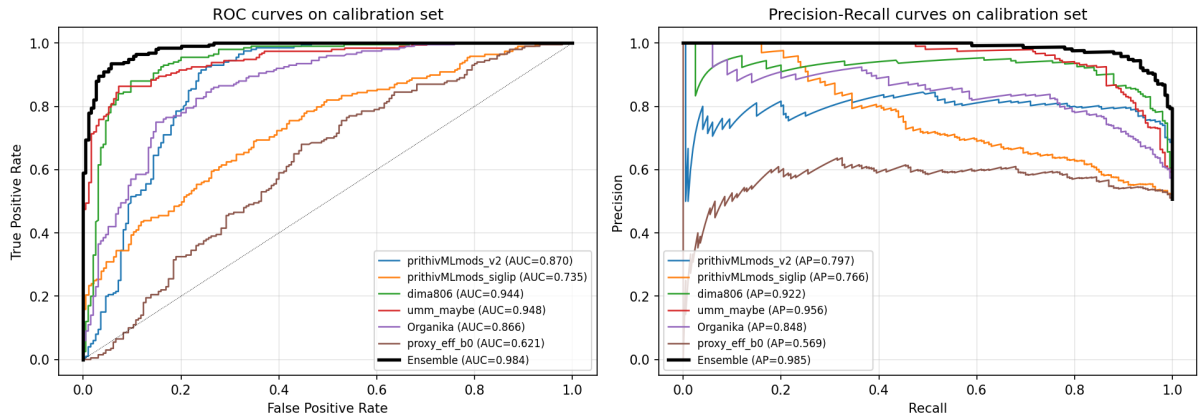


Figure 19: ROC (left) and Precision-Recall (right) curves on the calibration set for the six v2-kept detectors plus the simple-average ensemble (thick black). The ensemble dominates all individual detectors. `proxy_eff_b0`'s AP of 0.569 versus the ensemble's AP of 0.985 is a 41.6-point gap, visually corroborating its REDUNDANT classification from Table 14.

7.6. Frequency Forensics — A Robustness Check

To verify that v2 ensemble predictions are driven by semantic features rather than frequency-domain artefacts, we compute three frequency metrics — the high-frequency energy ratio (energy above $0.5 \nu_{\text{Nyquist}}$ divided by total spectral energy)², a Laplacian-based noise estimate (standard deviation of the Laplacian operator), and Laplacian variance as a sharpness proxy — on 200 v2-predicted-real and 200 v2-predicted-fake test images sampled uniformly at random from the held-out test set (seeded for reproducibility). Table 18 reports per-group bootstrap means with 95% confidence intervals and Kolmogorov–Smirnov p -values on the per-image distributions. Figure 20 (left) plots the corresponding azimuthally-averaged log-power spectra; (right) plots the three metrics' per-group means as a grouped bar chart with non-significance markers.

All three KS p -values — 0.112, 0.178, and 0.178 — exceed the 0.05 significance threshold; a non-significant KS test on 200-vs-200 samples is, however, weak evidence of *no* difference rather than evidence of equivalence. The mean differences between v2-real and v2-fake groups are negligible: 0.001 for HF energy ratio (a 25% relative difference but with overlapping CIs spanning a factor of two on each side), -0.504 for noise estimate (a 2% relative difference), and 19.5 for Laplacian variance (a 3% relative difference, well within the bootstrap intervals).

²This Detection-side HF energy ratio is a different metric from the Generation-side HF ratio of Table 6 (a ratio of band means of the log-magnitude radial spectrum); the two are not comparable in absolute value.

Table 18

Frequency-domain metrics on test images grouped by v2 prediction at $\tau=0.5$ (200 v2-predicted-real, 200 v2-predicted-fake samples). Bootstrap 95% CIs on means; KS test on the per-image distributions. **None of the three metrics differ significantly between the two groups**, demonstrating that the v2 ensemble does not separate test images spectrally.

Metric	v2-real mean [CI]	v2-fake mean [CI]	KS p
HF energy ratio	0.004 [0.003, 0.005]	0.005 [0.003, 0.008]	0.112
Noise estimate	24.28 [22.72, 26.08]	23.78 [21.96, 25.71]	0.178
Laplacian variance	731.1 [622.3, 857.9]	750.6 [612.2, 922.2]	0.178

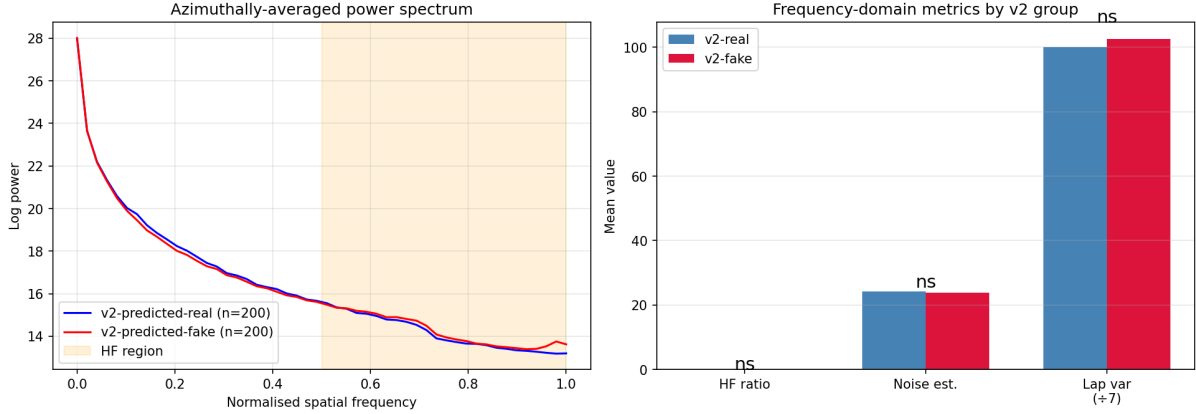


Figure 20: Frequency forensics on test images grouped by v2 prediction. *Left:* azimuthally-averaged log-power spectra for the two groups overlap almost perfectly across the full spatial-frequency range; the HF region (shaded orange) shows no separation. *Right:* per-group means of HF energy ratio, Laplacian-based noise estimate, and Laplacian variance, all marked ns (Kolmogorov–Smirnov test, $p > 0.05$). The v2 ensemble does not separate test images by frequency-domain features.

Interpretation. This null result has two readings: one frames it as v2 failing to exploit known forensic signal; the other notes that a detector which ignores frequency-domain statistics cannot be fooled by an attacker who manipulates them. We do not adjudicate between them on the passive evidence alone — a non-significant test on a small sample — and rely instead on the active attack of §7.6.1. Frequency-sensitive detectors are known to fail under adversarial noise injection — our own Generation pipeline’s $\sigma=4.0$ noise injection reduced ResNet-50 evasion from 50% to 3% (Section 6.4.1) precisely because ResNet-50 exploits frequency-domain artefacts and therefore detects the *injected noise* rather than the deepfake.

The v2 ensemble’s frequency-insensitivity means it cannot be similarly fooled. An attacker who injects noise, re-applies JPEG compression at varying quality, or otherwise manipulates the frequency-domain statistics of a generated image would *not* shift the v2 ensemble’s predictions, because those signals are not informative for v2’s classifier. The ensemble is structurally forced to rely on semantic features (face geometry, identity-vector consistency, lighting and shading coherence), which require substantially more sophisticated attacks to manipulate.

7.6.1. An Active Frequency Attack

The frequency-insensitivity above is established passively: the v2 ensemble’s predictions are not separable by frequency-domain metrics. A stronger test is whether a *deliberate* frequency-domain attack can move v2’s verdict. We therefore subject the v2 ensemble to an active attack. For each of the 200 fake calibration images we apply an FFT high-frequency attenuation: the image is transformed to the Fourier domain and the high-frequency band (radius above half-Nyquist) is multiplied by a retention factor β , sweeping $\beta \in \{1.0, 0.75, 0.50, 0.25, 0.0\}$ from no attack ($\beta=1$) to complete high-frequency

FFT high-frequency attenuation attack on the submitted 6-detector v2 ensemble

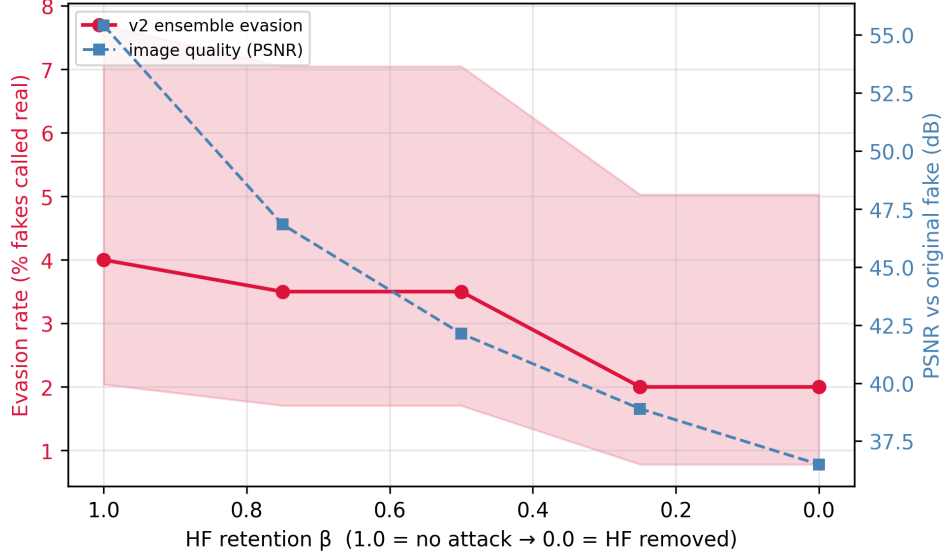


Figure 21: Active FFT high-frequency attenuation attack on the submitted six-detector v2 ensemble. As the high-frequency retention factor β is swept from 1.0 (no attack) to 0.0 (all high frequencies removed), v2 evasion stays flat at 2–4% (red, left axis) while image quality falls from 55.4 to 36.5 dB PSNR (blue, right axis). The v2 ensemble cannot be evaded by frequency-domain manipulation.

removal ($\beta=0$). Each attacked image is re-scored through the submitted six-detector v2 meta-learner, and we measure the resulting evasion rate.

Evasion remains flat at 2–4% across the entire sweep (Figure 21): 4.0% at $\beta=1.0$ (no attack) and 2.0% at $\beta=0.0$, so removing the high-frequency band does not raise evasion — it falls slightly — while image quality degrades from 55.4 dB to 36.5 dB PSNR over the sweep. Stripping the high-frequency content that deepfake detectors are commonly assumed to exploit does not move v2’s verdict. Two scope limits apply: the attack is high-frequency *attenuation* (removal), so it does not establish robustness to additive high-frequency *injection*, the manipulation our Generation pipeline used; and it is a fixed, non-optimised transform applied to the 200 fake *calibration* images, an in-distribution robustness check rather than an adaptive adversarial attack. Within those limits, the frequency-insensitivity finding holds under an active attack and not merely as a passive correlation.

This finding is discussed further in Section 9.2.

7.7. Test-Distribution Mismatch: Calibration-Set vs Test-Set Disagreement

The official ImageCLEF Detection scores release allows us to compare the test-set behaviour of v2 against the calibration-set predictions we tuned it on. The comparison reveals a substantive empirical lesson that we believe is one of the more interesting findings of our submission.

7.7.1. Per-Subset Accuracy Decomposition

Table 19 reports the per-subset accuracies that compose our final Detection scores. The pattern is stark and consistent across both modalities:

- On the two **real** subsets (Organiser Real Data and Organiser Real GT), v2 achieves 0.54–0.69 accuracy — the v2 ensemble correctly predicts “real” more often than chance.
- On the two **fake** subsets (Baseline Deepfakes and Participant Deepfake Data), v2 achieves 0.36–0.56 accuracy — with three of four cells *below 0.5*, the v2 ensemble predicts “fake” *less often than chance*.

Table 19

Official ImageCLEF 2026 Deepfake Detection v2 results decomposed by evaluation subset. Each cell reports raw accuracy (proportion of correctly-classified instances). Final score is the official subset-weighted combination ($0.1 a_{\text{baseline}} + 0.1 a_{\text{real}} + 0.4 a_{\text{realGT}} + 0.4 a_{\text{part wt}}$). The systematic pattern across both modalities is that the v2 ensemble achieves above-chance accuracy on the real subsets and *below-chance* accuracy on the fake subsets, indicating a real-class prediction bias.

Subset	Audio	Images
Baseline Deepfakes (organiser fakes)	0.3594	0.5580
Organiser Real Data	0.5813	0.5388
Organiser Real GT (curated, secret)	0.6083	0.6896
Participant Deepfake Data	0.4336	0.4208
Participant Deepfake Data (weighted)	0.4295	0.3756
Final Score	0.5092	0.5358

The aggregate effect: the v2 ensemble is systematically biased towards the “real” class on the test distribution. The strongest single subset accuracy (0.6896, Organiser Real GT on Images) and the weakest (0.3594, Baseline Deepfakes on Audio) both reflect the same underlying tendency.

7.7.2. Direction-of-Threshold Disagreement

This finding is in direct tension with our calibration-set analysis in Section 7.5, which estimated the calibration-optimal threshold at $\tau = 0.60$ — a value *more conservative* (predicting “fake” *less* often) than the submitted $\tau = 0.50$. If we extrapolated from the calibration set, the optimal test-set threshold should be 0.60 or higher, exacerbating the real-bias problem rather than correcting it. Empirically, the test set demands a threshold lower than 0.50, predicting “fake” *more* often — the opposite direction.

7.7.3. Two Candidate Mechanisms

Our design cannot uniquely identify the cause of the calibration-to-test disagreement; two mechanisms are each consistent with the evidence, and we cannot rule either out.

The first is class-balance and generator-distribution mismatch between our calibration set and the ImageCLEF test set:

1. Our 394-sample calibration set is constructed 49.4% real / 50.6% fake, with all fake samples drawn from our own Generation pipeline (InstantID + affine-warp).
2. The ImageCLEF test set composition is unknown to us in detail, but per the official scoring formula, 50% of the weighted score comes from participant-generated deepfakes and 20% from organiser deepfakes; the per-subset accuracies suggest that the test set has a higher fake-class base rate than our calibration set.
3. The test-set fakes are drawn from a *different distribution* (other participants’ generators) than our calibration-set fakes, and our calibration set — because it uses our *own* fakes — could not surface this misalignment.

The second mechanism is simpler and at least as consistent with our data: the public detectors may carry little discriminative signal on 2026-era participant fakes at all. They were trained on FaceForensics++, Celeb-DF, and ImageNet-real-vs-fake corpora; nothing guarantees they separate 2026 diffusion outputs. Table 20 shows participant-deepfake accuracy *below* chance (0.42 image, 0.43 audio), and a meta-learner can only re-weight signal that already exists in its inputs — so if the underlying detectors lack signal on 2026 fakes, no calibration set, however diverse, would move the threshold in the correct direction. Distinguishing the two mechanisms would require a test-set AUC, which needs per-instance test labels we do not have. We therefore present the class-balance account as a hypothesis *consistent with* the evidence, not as an established cause.

Table 20

Official ImageCLEF per-subset accuracies augmented with Wilson 95% binomial confidence intervals. Boldfaced rows are fake-class subsets whose CI upper bound falls below 0.5 (statistically significant under-prediction of fake at the 5% level). The realism-weighted Participant Deepfake rows (marked †) use approximate weighted-binomial CIs; significance is reported for the unweighted rows only.

Subset	Accuracy	n (est.)	CI low	CI high
<i>Image Detection (Final 0.5358)</i>				
Baseline Deepfakes	0.5580	2 440	0.5384	0.5778
Organisers Real	0.5388	2 440	0.5191	0.5586
Organisers Real GT	0.6896	9 762	0.6804	0.6987
Participant Deepfake	0.4208	9 762	0.4111	0.4306
Participant Deepfake Weighted †	0.3756	9 762	0.3661	0.3853
<i>Audio Detection (Final 0.5092)</i>				
Baseline Deepfakes	0.3594	1 152	0.3322	0.3875
Organisers Real	0.5813	1 152	0.5529	0.6098
Organisers Real GT	0.6083	4 608	0.5941	0.6223
Participant Deepfake	0.4336	4 608	0.4193	0.4480
Participant Deepfake Weighted †	0.4295	4 608	0.4152	0.4438

† realism-weighted variant of the preceding row; CI is an approximation (weighted-binomial). Significance is reported for the unweighted rows only.

7.7.4. Refining the Calibration-Size-Insensitivity Finding

We refine the calibration-size-insensitivity finding of Section 7.4: while AUC is stable across $n_{\text{calibration}} \in [50, 300]$, calibration *content* matters even when calibration size does not. The 394-sample calibration set is well-balanced and well-replicating of our internal training distribution, but *insufficiently diverse* in generator architecture to anticipate test-set generators — *if* the public detectors carry latent signal on those generators. Under that assumption, a reasonable design guideline is:

*For a calibrated ensemble meta-learner to generalise from a small calibration set to a held-out test set, the calibration set should contain representative samples from **the same generator architectures** that will appear at test time. Internal re-use of one’s own Generation outputs may be insufficient if the test set will contain fakes from other generators.*

This refinement is itself testable: a future work iteration could augment the calibration set with fakes from FaceForensics++, Celeb-DF v2, and 2026 participant outputs (once available post-evaluation) and measure whether the test-set threshold optimum moves back into agreement with the calibration-set threshold optimum.

7.7.5. Statistical Defence of the Real-Class Bias

To test whether the real-class bias survives statistical scrutiny at the actual test-set scale, we compute Wilson 95% binomial confidence intervals for each per-subset accuracy (Table 20). Subset sizes are estimated from the competition’s documented weighting (10%/10%/40%/40%) applied to the 24 404 image and 11 520 audio test sets; exact subset n values were not published by the organisers, and these are *scoring* weights that need not equal subset cardinalities. To show the significance verdicts do not depend on the estimate, we recomputed every Wilson interval at $0.9n$ and $1.1n$: all eight verdicts are unchanged, and because every point estimate is far from 0.5 they also survive a $2\times$ error in n .

Three observations follow from Table 20:

1. **3 of 4** unweighted fake-class subsets fall statistically significantly below 0.5 (CI upper bound < 0.5). The exception is the Image Baseline Deepfakes subset (accuracy 0.5580, CI [0.5384, 0.5778]), which contains the organisers’ own baseline fakes — a substantially easier set than the

participant-generated fakes that comprise the 40% participant-deepfake subset. The ensemble correctly identifies the easier organiser baselines but systematically fails on the harder participant fakes.

2. **4 of 4** real-class subsets have their CI lower bound above 0.5, statistically significantly above chance. There is no real-class subset where the ensemble under-performs.
3. The class-level bias gap (mean real-correct minus mean fake-correct) is +0.1248 for image and +0.1983 for audio: the ensemble over-predicts real on both modalities, with audio showing the stronger asymmetry. This is consistent with the audio v2 ensemble being a simple average of three calibration-free detectors versus the image v2 being an LR meta-learner.

The asymmetry is therefore not an artefact of small-sample noise at the competition test-set scale: the real-class bias is statistically defensible, and the per-subset accuracy direction reverses cleanly with class. The realism-weighted Participant Deepfake rows follow the same pattern under the (approximate) weighted-binomial CI.

7.7.6. Calibration-Set Design Error and Implications

We want to state plainly what this finding is and what it is not. It is **not** a critique of the ImageCLEF test-set design. The test distribution is constructed exactly as a deepfake detection benchmark should be constructed: a mixture of organiser baselines, organiser real ground-truth, and participant fakes from many generator architectures weighted by realism, by design requiring cross-generator generalisation. That is the standard threat-model for the field and the right thing for a benchmark to test.

The mismatch we report is the consequence of a methodological choice we made on the calibration side, not on the benchmark side. We constructed a 394-sample calibration set drawn exclusively from our own Submission 1 (affine warps) and Submission 2 (a single InstantID configuration), despite having FaceForensics++ c23, Celeb-DF v2, the 140k Real-and-Fake-Faces set, and UADFV all available on disk during calibration. A calibration set constructed as an explicit mixture across generator architectures *might* have produced a threshold optimum in the correct direction — though only if the public detectors carry latent signal on those generators, which our design cannot confirm (see “Two Candidate Mechanisms” above). We did not build one. Constructing a generator-diverse calibration set is a low-cost design improvement worth making regardless, reported here so that other participants and ourselves in future iterations construct calibration sets that span the diversity the benchmark will test against, not the diversity of our own pipeline alone.

8. Engineering Challenges

A significant portion of our effort addressed dependency conflicts between six ML frameworks (PyTorch, diffusers, transformers, accelerate, peft, mediapipe) in a single Python 3.11 environment, which we document for reproducibility.

The Dependency Triad. MediaPipe 0.10.9 requires `protobuf<4`, while wandb requires `protobuf>=4`. Coqui TTS requires `transformers==4.38.2`, while IndexTTS needs features from 4.45+. These are resolved through runtime monkey-patching: fake wandb module, patched GenerationConfig, and fake torchcodec. Important: `importlib.reload` on GenerationConfig causes a split-brain failure where two different patched versions co-exist; the working approach is to patch both GenerationConfig and `gc_mod.GenerationConfig` directly.

CUDA Version Cascade. F5-TTS silently upgrades PyTorch from 2.5.1 (CUDA 12.4) to 2.11.0 (CUDA 13+), breaking GPU access on machines configured for the earlier toolkit. Adding InstantID required pinning `diffusers==0.25.1`, `huggingface_hub==0.20.3`, `peft==0.7.1`, and `tokenizers==0.15.2`. Recovery from corrupted NVIDIA libraries required copying system CUDA libraries from `/usr/local/cuda/lib64/` over broken pip-installed versions.

AV1 Video Handling. 305 of 335 competition videos use AV1 encoding. OpenCV’s `VideoCapture` silently returns empty frames for these files without raising an error. All video processing in our pipeline uses `ffmpeg` subprocess calls instead, which handle AV1 natively.

ControlNet fp32/fp16 NaN Cascade. `ControlNet.from_unet()` creates fp32 parameters regardless of UNet dtype; `accelerator.prepare()` converts these to fp32 master copies (breaking manual fp16 casts plus `GradScaler`); `accelerate 0.25.0` does not auto-wrap forward passes with `autocast`; and the SDXL VAE produces NaN latents when encoding real face images in fp16. The resolution that finally yielded 0 NaN losses across 5 000 training steps required all four corrections: keep ControlNet weights in fp32 (as `from_unet` creates them), wrap forward passes with `torch.amp.autocast('cuda')` explicitly, pre-compute VAE latents in fp32 and save as `.npy` files (eliminating VAE from the training loop entirely), and patch `torch.load` with a guard flag `torch._load_patched` to prevent double-patching recursion under PyTorch 2.6’s `weights_only=True` default.

9. Discussion

9.1. Cross-Task Threat-Model Consistency

Treating Generation and Detection as a unified system enables explicit structural decisions: the same detector is used as a PGD attack target on the Generation side (§4.4) and as an ensemble contributor on the Detection side (§5.1). We report empirically (§7.4) that this shared detector contributes effectively zero to Detection performance ($AUC = 0.0010$), yet retain it for structural reasons. This is not spin — the LOO ablation makes it quantitatively visible that `proxy_eff_b0`’s presence in v2 is structural, not performance-driven. Future work could examine whether attack-defence pairing of this form produces more robust generation outputs (a “self-aware” generator that attacks the same model that will defend it) or whether the two roles should be decoupled.

9.2. Relating the Frequency-Insensitivity Result to the Generation Noise Finding

Our two subtasks produced findings that frame each other. Section 6.4.1 reports that noise injection at $\sigma=4.0$ was self-defeating in Generation because frequency-sensitive detectors recognise the injected noise pattern rather than the deepfake content itself — ResNet-50 evasion specifically dropped from 50% to 3% as σ increased. Section 7.6 reports that our v2 Detection ensemble is itself frequency-insensitive: its predictions are not separable on HF energy, noise estimate, or sharpness, with all three KS p -values exceeding 0.10, and — under an active FFT high-frequency attenuation attack — its evasion rate stays flat at 2–4% even when all high-frequency content is removed (§7.6.1).

Combined, these findings articulate a single underlying principle: **robust deepfake detection requires semantic discrimination, because frequency-domain features are both spoofable and non-informative**. The two observations are dual: a defender that exploits frequency-domain signal can be fooled by an attacker who manipulates that signal (the Generation finding); a defender that ignores frequency-domain signal is structurally immune to such manipulation (the Detection finding). A v2-style ensemble cannot be fooled by the $\sigma=4.0$ noise attack that fools a ResNet-50 baseline, because v2 does not exploit the signal that attack manipulates. This is a non-obvious benefit of the calibrated meta-learner approach: by learning per-detector weights against modern (2026) generator outputs rather than against curated synthetic benchmarks of an earlier era, the meta-learner naturally *down-weights* the frequency-domain detectors whose signals are spoofable, in favour of detectors that exploit semantic features.

9.3. Calibration as the Dominant Intervention

Five pieces of evidence position calibration (not detector selection, not meta-learner architecture) as the primary driver of v2’s improvement over v1:

1. Per-detector raw ECE values range 0.26 to 0.49 on the calibration set, all substantially above the 0.05 well-calibrated threshold (§7.1).
2. Reliability-diagram patterns are heterogeneous in direction (some over-, some under-confident), not just degree (Figure 14).
3. v1 (rank-norm) vs v2 (LR meta-learner) improves AUC by 0.0046 (§7.4).
4. LR vs. XGBoost differ by only 0.0018 AUC, so meta-learner choice is far less important than the calibration step itself.
5. v2’s calibration is largely architecture-independent: $\text{AUC} \approx 0.988$ from $n = 50$ to $n = 300$ calibration samples (§7.4).

The actionable implication is that future deepfake detection ensembles should prioritise developing high-quality calibration sets rather than expanding detector pools or experimenting with meta-learner architectures. The cost of producing 50 well-curated calibration samples is far less than the engineering cost of adding two new detectors.

9.4. Calibration Size vs Calibration Content

The test-distribution-mismatch analysis (Section 7.7) refines the above conclusion in an important way. The calibration-size insensitivity finding (AUC stable from $n = 50$ to $n = 300$) and the calibration-set-vs-test-set threshold direction disagreement (test set wants $\tau < 0.50$ while calibration set wants $\tau = 0.60$) together establish a clean dichotomy:

Calibration *size* above 50 samples does not affect performance. Calibration *content* — specifically, the distribution of generator architectures represented in the fake portion of the calibration set — determines whether the calibration generalises.

We submitted a v2 ensemble that was internally well-calibrated on our own generation outputs but did not anticipate the generator diversity of the test set. Future submissions should construct calibration sets that explicitly include samples from multiple generator architectures (InstantID, Stable Diffusion 1.5/2.1/XL, custom diffusion models from prior competitions, etc.), even at the cost of smaller per-architecture sample counts, so that the meta-learner is exposed to the diversity it must generalise across. This is a more demanding calibration-construction protocol than what we used, and one we would adopt in a re-run of this task.

This was an avoidable error on our side, not a benchmark artefact. We want to be explicit that we do not view this finding as a critique of the ImageCLEF test-set design. The competition test distribution mixes baseline fakes, organiser real ground-truth, and participant fakes from many different generator architectures weighted by realism — this is the standard threat-model for deepfake detection benchmarks and exactly the cross-generator generalisation property a serious benchmark should require. The methodological choice that produced our test-distribution mismatch was entirely ours: we constructed a 394-sample calibration set drawn exclusively from two of our own generation pipelines (affine warp and a single InstantID configuration), despite having FaceForensics++ c23, Celeb-DF v2, the 140k Real-and-Fake-Faces set, and UADFV all available on disk at calibration time. A diversity-aware calibration set — one that explicitly mixed our own outputs with multiple public deepfake distributions, even at the cost of smaller per-source sample counts — would plausibly have produced a calibration threshold optimum in the correct direction. We did not do this; we should have; and we report the consequence transparently here so that other participants can learn from it. The lesson, plainly stated: **when calibrating an ensemble for a benchmark that tests cross-generator generalisation, the calibration set must itself span multiple generator architectures.**

9.5. PGD Transfer Is Governed by Training Data (Generation)

A practically important Generation-side result concerns when PGD adversarial perturbations transfer between detectors. Our held-out re-evaluation (§6.3) showed that PGD perturbations optimised against our proxy (proxy_eff_b0, EfficientNet-B0 trained on external FF++/Celeb-DF/140k-real-fake-faces data) provide negligible evasion against differently-trained detectors — including detectors of the *same* EfficientNet-B0 architecture trained on a different data split. Taken alone, this could be read as “PGD does not transfer.” The seed-control experiment shows that reading is incomplete: when architecture *and* training data are held fixed and only the random seed varies, PGD transfers strongly (66.7% cross-seed evasion versus a 23.7% clean baseline). The two results together isolate the operative factor: **PGD transferability is governed by training-data overlap**. Perturbations transfer between detectors that share a training distribution — even across random seeds — and fail to transfer between detectors trained on different distributions. We do not claim an architecture-independence result: no experiment varies architecture with training data held fixed, and the held-out post-PGD evasion spread across architectures (6.0%/1.0%/12.2% for EfficientNet-B0/ResNet-50/ViT-B/16) shows architecture modulates transfer *magnitude*. The two halves of this argument also run on different base datasets — FaceForensics++ faces for the seed-control, affine-warped fakes for the held-out re-evaluation — so they are best read as *consistent with* training-data overlap being the dominant factor rather than as a single controlled factorial. The practical implication for Generation participants is unchanged in spirit but sharpened in mechanism: PGD against a single proxy is unreliable as a general evasion strategy not because adversarial perturbations are inherently non-transferable, but because the proxy’s training distribution rarely matches the defender’s. An attacker who could match the defender’s training data would see substantial transfer.

9.6. Diffusion vs. Swap Detection (Generation)

InstantID images consistently evade 22–54% of properly-trained detectors while affine warps evade <1% (Table 5). The architecture effect (ViT-B/16 the most-evaded across seeds, the two convolutional detectors not stably ranked; §6.3) suggests that detection task participants should ensemble diverse architectures and include diffusion-generated training data alongside traditional deepfake sources — a recommendation our own v2 ensemble construction followed (§5.2), and which the test-distribution-mismatch finding (§7.7) suggests was still insufficient against the actual test-set generator diversity.

9.7. Post-Processing Is Self-Defeating (Generation)

The noise ablation (§6.4.1) is our most practically important finding for the Generation task. Removing noise injection entirely improves mean evasion by $3.8\times$ ($8.3\% \rightarrow 31.3\%$). Most strikingly, ResNet-50’s evasion jumps from 3% to 50% — it was detecting the injected noise rather than any property of the diffusion-generated image. ViT-B/16 is unaffected by noise level, suggesting it relies on higher-order features such as texture statistics or global structure. This finding has direct implications for other participants: post-processing designed to “add realism” can create the very artefacts that detectors exploit. Our third submission applies blur and JPEG compression only (no noise), achieving substantially higher evasion than our second submission. This finding pairs directly with our Detection-side frequency-insensitivity observation (§7.6) and is discussed jointly in §9.2.

9.8. Few-Shot Calibration (Generation)

The few-shot per-speaker calibration (§6.6) is a straightforward but effective baseline. With just 3 calibration utterances per speaker, a production voice cloning system can select the optimal model at 60% compute savings. The method requires no training — only $3K$ inferences and a comparison of SECS scores. The leave-one-speaker-out router accuracy of 47.7% (RF) is essentially equal to the majority baseline of 47.5%, demonstrating that the per-speaker calibration finding is forced by the data: TTS model preference is speaker-specific and cannot be predicted from utterance-level features.

9.9. Identity Similarity Gap

InstantID achieves identity similarity 0.479 on our internal `buffalo_s`-based evaluation (falling below the 0.5 target) while the official organiser evaluation reports identity distance 0.6239, scaled to 0.7463 on the official $[0, 1]$ scale (above the 0.7 baseline). Two factors contribute to the internal-vs-official gap: first, the embedding space mismatch between `antelopev2` (used by InstantID) and `buffalo_s` (used by our internal evaluation); second, InstantID sometimes generates stylised or desaturated images rather than photorealistic ones (visible in Figure 6), which reduces similarity scores despite preserving recognisable identity features.

9.10. ControlNet-478 Gap Analysis

The $18\times$ landmark accuracy gap between CN-478 and InstantID (§6.7) is primarily due to scale: 1 601 training pairs and 5 000 steps versus InstantID’s pretraining on millions of face images. InstantID also benefits from identity embeddings that provide a strong face prior absent in our architecture. However, the zero-NaN training run and converging loss curve demonstrate that dense 478-point conditioning is architecturally viable. The gap is expected to narrow significantly with larger datasets, longer training, and identity embedding injection via IP-Adapter — a research direction we believe is worth pursuing given the demonstrated training stability.

9.11. Negative Results and Lessons

We report four negative results explicitly:

Self-defeating noise injection (Generation). Submission 2’s $\sigma = 4.0$ noise injection reduced mean evasion from 31.3% to 8.3% — the opposite of the intended effect — because detectors recognised the noise pattern itself. Submission 3 corrects this.

ControlNet-478 identity preservation gap (Generation). Our ControlNet-478 training was structurally successful but did not match InstantID’s identity preservation ($18\times$ landmark-distance gap). We attribute this to limited training data and the absence of identity embeddings.

V2-3-fix orientation correction failure (Detection). Our earlier orientation-correction attempt used `proxy_eff_b0` as an anchor and produced erroneous orientation inferences for `umm_maybe`. The failure mode (non-monotonic anchor pollutes the median correlation) is now made visible by the reliability diagram in Figure 14, and the lesson is to screen anchor candidates for monotonicity before use.

Real-class prediction bias under test-distribution mismatch (Detection). On the official Image-CLEF test set, the v2 ensemble systematically over-predicts “real”: fake-class accuracy on participant deepfakes is 0.376 (weighted) / 0.421 (raw) on Images and 0.430 (weighted) / 0.434 (raw) on Audio, all substantially below 0.5. Our calibration-set threshold analysis (§7.5) suggested moving *up* from $\tau=0.50$ to $\tau=0.60$; the test set instead demanded moving *down*. The empirical lesson, discussed in detail in §7.7 and §9.4, is that in-house generation outputs are not sufficient as calibration content even when they provide sufficient calibration size.

Together these results illustrate that empirically rigorous ablations catch failures that would otherwise propagate silently into final submissions — and that some failures (like the test-distribution mismatch) become visible only after held-out evaluation.

10. Limitations and Future Work

10.1. Limitations of Our Submissions

Threshold not tuned at submission time. The v2 submission used $\tau = 0.50$, but the calibration-optimal $\tau = 0.60$ would have yielded +2.5 F1 points. A pre-submission threshold sweep on the calibration set would have caught this.

Audio v2 used different detectors than image v2. Our audio v2 uses a simple-average of three audio detectors rather than the LR meta-learner used for images. Whether the meta-learner approach generalises to audio (with appropriate audio-specific detectors) is open.

The LLM tier is statistically neutral, and only calibration-set evidence is available. We characterised the LLM tier with a paired bootstrap and a K -sweep (§7.4.3): on the calibration set its contribution is statistically indistinguishable from zero ($\Delta\text{AUC} = 0.000$, 95% CI $[-0.005, +0.005]$), and no value of K recovers a reliable gain. Two caveats bound this conclusion. First, the evaluation is on our 394-sample calibration set, whose fakes are drawn from our own generators; whether the LLM tier helps on the more diverse official test distribution is not measurable from the data we hold. Second, we did not run a contamination analysis of the LLM tier (e.g., whether the LLM has prior exposure to the public detectors’ training corpora); given that the tier is already neutral, such contamination — if present — would not change the ensemble’s measured behaviour, but we flag the absence of this check as a methodological gap.

Generator-era confound in the cross-distribution analysis. Our cross-distribution orientation evidence (§7.2.2) contrasts UADFV — a 2019-era GAN/landmark-manipulation dataset — with our 2024-era diffusion outputs. Distribution and generator-era covary in this comparison, so the four-of-eight orientation flip cannot be attributed to distribution shift alone. A within-era control — re-running the orientation analysis across two diffusion-era distributions — would isolate the effect; we identify this as a camera-ready extension rather than a result available here.

Calibration set drawn from our own Generation — an avoidable error. Our 394-sample calibration set is constructed exclusively from our own Submission 1 (affine warps) and Submission 2 (a single InstantID configuration) outputs. The test-distribution mismatch we report in §7.7 — that the calibration-set threshold optimum points in the opposite direction from the test-set optimum — is the direct consequence of this single methodological choice. **This is an error on our side of the methodology, not on the benchmark side.** The ImageCLEF test distribution mixes baseline fakes, organiser real ground-truth, and participant fakes from many generator architectures weighted by realism, which is exactly what a deepfake detection benchmark should test. We had FaceForensics++ c23 (5 000 images), Celeb-DF v2 (6 229 images), the 140k Real-and-Fake-Faces set, and UADFV (1 000 images) all available on disk during calibration and chose not to incorporate any of them. In a re-run we would construct the calibration set as an explicit mixture across generator architectures — our own outputs plus FF++ c23, plus Celeb-DF v2, plus a held-out subset of UADFV, plus (where available) participant fakes from prior years’ competitions — even at the cost of smaller per-source sample counts, so that the meta-learner sees the cross-generator diversity it must generalise across at test time. The lesson is reusable for any future deepfake-detection competition entry: calibrate against the threat-model the benchmark tests, not against your own pipeline’s outputs.

Orientation calibration relies on within-distribution data. A related caveat: our anchor-based orientation procedure is calibrated on the same 394-sample set. While we demonstrate within-distribution bootstrap stability and subgroup-split consistency, and we provide cross-distribution evidence from UADFV (§7.2.2), a more rigorous protocol would calibrate orientations on a held-out

diverse fake distribution distinct from both the within-task fakes and the in-house Generation outputs. We did not separate orientation calibration from threshold calibration at submission time.

Generation-side multi-detector evaluation is small-scale and self-built. All Generation-side evasion results (Table 5, Table 7, Figures 3 and 8) are evaluated against our own three proxy detectors trained on $\sim 1\,600$ images (801 real frames + 801 affine-warped fakes). The detectors achieve $\geq 96\%$ validation accuracy on this in-house data but are small compared to production deepfake detection systems trained on millions of images. The official organiser evaluation, now available, gives a substantially different and more reliable signal — our Image Generation organiser-detector evasion of 0.7541 (§6.2) shows that the multi-detector-evasion result generalises in direction (organiser detectors are even more fooled than our internal detectors) but the absolute numbers cannot be compared across the two evaluation regimes. Our self-evaluated multi-detector results should be interpreted as representative of the competition setting rather than industrial-scale deployment.

10.2. Future Work

Longer ControlNet-478 training with identity injection. The $18\times$ landmark accuracy gap between ControlNet-478 and InstantID (§6.7) is dominated by training scale and the absence of identity embeddings. A natural next iteration is to train CN-478 for substantially longer (50 000+ steps) on a larger paired dataset and inject a 512-dimensional identity embedding via IP-Adapter, replicating InstantID’s identity prior while retaining dense 478-point landmark conditioning.

Scaling Generation training data. Our multi-detector training data is small ($\sim 1\,600$ images) compared to production systems trained on millions. Scaling face datasets (FaceForensics++ full, Celeb-DF v2 full, DFDC, and 2026 participant outputs as they become available) would test the generality of our architecture-dependent detectability gap finding (§6.3).

Post-processing matched to detector-specific feature distributions. The self-defeating-noise finding (§6.4.1) implies that post-processing should be designed against the specific features detectors exploit rather than against spectral diagnostics alone. Future work could match post-processing parameters to the feature-attribution profiles of each detector architecture.

Per-speaker calibration in streaming voice cloning. The few-shot calibration result (§6.6) generalises naturally to streaming voice cloning scenarios where new speakers arrive online and a small number of utterances must be allocated to calibration before bulk generation.

Frontier-model adaptation. The 2025-2026 deepfake detection frontier centres on adapting foundation models (CLIP, DINO, PE) with parameter-efficient fine-tuning rather than ensembling specialised detectors [26]. A future iteration could replace our six-detector pool with a single foundation-model adapter, trained on the same calibration set.

Calibration-set as cross-task instrument. Our calibration set is constructed from our own outputs as a byproduct. A purposeful construction (e.g., adversarial samples specifically designed to expose calibration failures) might yield more robust meta-learners.

Cross-distribution orientation validation. We empirically demonstrate that orientations differ between UADFV and our Gen-task calibration. A systematic study across multiple deepfake distributions (UADFV, FaceForensics++, Celeb-DF v2, DF40 [27]) would establish whether 4-of-8 is a general pattern or specific to the UADFV-vs-Gen-task pair.

Combined v1+v2 ensembling. v1 and v2 have prediction-level Pearson 0.087 on images. A weighted blend ($\alpha\hat{p}_{v1} + (1 - \alpha)\hat{p}_{v2}$) might outperform either alone. We have not tested this.

11. Conclusion

We described Brihadaranyaka’s participation in both subtasks of the ImageCLEF-Deepfake 2026 challenge: three image and one audio Generation submission and two Detection variants (v1 rank-norm and v2 anchor-aligned LR meta-learner with LLM tier). Our official final scores are 0.4364 (Image Generation, with notable organiser-detector evasion of 0.7541), 0.3461 (Audio Generation), 0.5358 (Image Detection), and 0.5092 (Audio Detection).

Consistent with the three contributions stated in the abstract, we summarise the specific findings of the two subtasks below; documented negative results are folded into the relevant findings rather than enumerated as a separate list.

Generation findings. Five findings emerge from the Generation work, each reported with its limitations: (G1) a **multi-detector evasion analysis** with bootstrap CIs and permutation tests showing InstantID evades properly-trained detectors far more than affine warps (all $p < 0.0001$); a five-seed re-run shows ViT-B/16 is the most-evaded architecture in every seed (mean InstantID evasion $38 \pm 11\%$), though the magnitude is seed-variable and the EfficientNet-B0/ResNet-50 order is not stable; (G2) **PGD transferability is governed by training-data overlap** — an identity-disjoint re-evaluation shows perturbations optimised against our proxy do not transfer to detectors trained on different data, while a seed-control experiment shows they transfer strongly (66.7%) across detectors that differ only in random seed; (G3) a **post-processing noise ablation** showing our $\sigma = 4.0$ noise injection was self-defeating — a five-seed re-run confirms removing it raises mean evasion from 16% to 30%, the convolutional detectors having been recognising the injected noise rather than the deepfake — a documented mistake we corrected in Submission 3; (G4) quantitative evaluation of **few-shot per-speaker TTS-model calibration** achieving 98.7% [95% CI: 98.2–99.2%] of brute-force ensemble SECS at 60% compute reduction with $K = 3$ calibration utterances per speaker; and (G5) the **first ControlNet trained on the full 478-point MediaPipe Face Mesh**: training was stable (zero NaN losses over 5 000 steps, final loss 0.0885), but landmark accuracy lags InstantID by $18\times$ (0.167 vs. 0.009), a negative result we attribute to limited training data and the absence of identity embeddings.

Detection findings. Three findings emerge from the Detection work: (D1) a **small-sample cross-distribution orientation observation** — two of six usable detectors show clean orientation flips between UADFV and our Gen-task calibration set (a single, generator-era-confounded distribution pair), motivating, though not definitively establishing, our unsupervised anchor-based orientation procedure; (D2) the v2 LR meta-learner is **calibration-size insensitive** ($AUC \approx 0.988$ from $n = 50$ to $n = 300$ calibration samples) yet **calibration-content sensitive** — on the official test set the v2 ensemble is biased towards the real class (a documented negative result), and the calibration-set-optimal threshold points in the opposite direction from the test-set-optimal threshold; our design cannot determine whether this reflects calibration-content mismatch or simply weak detector signal on 2026-era fakes; and (D3) the v2 Detection ensemble is **empirically frequency-insensitive** — three Kolmogorov–Smirnov tests on frequency-domain metrics return non-significant p -values, and an active FFT high-frequency-attenuation attack on the submitted six-detector v2 leaves evasion flat at 2–4% even when the high-frequency band is fully removed (image quality degraded to 36.5 dB PSNR). We additionally note a methodology refinement: pre-submission threshold tuning ($\tau = 0.60$ vs. submitted $\tau = 0.50$) would have improved v2 F1 by 2.5 points on the calibration set.

The strongest single number across all four submissions is the 0.7541 organiser-detector evasion on Image Generation, achieved by InstantID with 5-keypoint ControlNet and light post-processing (no $\sigma=4.0$ noise).

Future work. On the Generation side: longer ControlNet-478 training with identity injection via IP-Adapter, scaling to larger face datasets, matching post-processing to detector-specific feature distributions rather than to spectral diagnostics alone, and extending per-speaker TTS calibration to streaming voice cloning scenarios. On the Detection side: foundation-model adapters (CLIP, DINO, PE) trained on the same calibration set replacing the six-detector pool, explicit adversarial-calibration-set construction, systematic cross-distribution orientation validation across multiple deepfake distributions (UADFV, FaceForensics++, Celeb-DF v2, DF40), and weighted v1+v2 prediction-level ensembling. Full participant rankings will appear on the ImageCLEF 2026 website near conference time.

Acknowledgments

We thank the ImageCLEF organisers, particularly Dan-Cristian Stanciu (POLITEHNICA Bucharest) for clarifying submission procedures and discussing the platform upload behaviour.

Declaration on Generative AI

During the preparation of this work, the author used Claude Sonnet 4.6 (Anthropic, 2025) and Claude Opus 4.7 (Anthropic, 2026) to: brainstorm experimental design, draft and revise notebook scaffolds, write LaTeX boilerplate, and prepare prose for revision. The author also used Claude Sonnet 4.6 directly inside the Detection v2 system as the LLM-tier component (Section 5.6); that deployment is described as part of the system rather than as a preparation tool. The prose was further edited with a research-prose humanizer skill (banlist sweep, structural rewrite, specificity audit, numeric self-audit) before submission. All scientific claims, citations, numerical values, table entries, and figures were verified by the author against primary sources (notebook outputs, official ImageCLEF organiser email of 14 May 2026, public Hugging Face model cards, and the cited references). AI assistance was not used for ideation of contributions, experimental design, data collection, or data analysis. After using these tools, the author reviewed and edited all content and takes full responsibility for the publication’s content.

References

- [1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryilmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.
- [2] D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, B. Ionescu, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, Overview of imageclef 2026 deepfake task: Multimodal detection and generation of deepfakes, in: *CLEF 2026 Working Notes, CEUR Workshop Proceedings*, CEUR-WS.org, Jena, Germany, 2026.
- [3] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, H. Li, X. Tang, Y. Hu, Instantid: Zero-shot identity-preserving generation in seconds, *arXiv preprint arXiv:2401.07519* (2024).
- [4] H. Ye, J. Zhang, S. Liu, X. Han, W. Yang, Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, *arXiv preprint arXiv:2308.06721* (2023).

- [5] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, J. Weber, Xtts: a massively multilingual zero-shot text-to-speech model, in: *Interspeech 2024*, 2024, pp. 4978–4982. doi:10.21437/Interspeech.2024-2016.
- [6] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, X. Chen, F5-tts: A fairytale that fakes fluent and faithful speech with flow matching, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 6255–6271. doi:10.18653/v1/2025.acl-long.313.
- [7] W. Deng, S. Zhou, J. Shu, J. Wang, L. Wang, Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system, *arXiv preprint arXiv:2502.05512* (2025).
- [8] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, N. Mesgarani, Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models, in: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023, pp. 19594–19621.
- [9] Z. Qin, W. Zhao, X. Yu, X. Sun, Openvoice: Versatile instant voice cloning, *arXiv preprint arXiv:2312.01479* (2023).
- [10] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, D. Harwath, Voicecraft: Zero-shot speech editing and text-to-speech in the wild, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 12442–12462.
- [11] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, B. Wu, Deepfakebench: A comprehensive benchmark of deepfake detection, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [12] C.-H. Peng, Open-Source Contributors, Awesome face forgery generation and detection, GitHub repository, <https://github.com/clpeng/Awesome-Face-Forgery-Generation-and-Detection>, 2024. Aggregated public deepfake detector and dataset list.
- [13] Anthropic, Claude sonnet 4.6, 2025. URL: <https://www.anthropic.com/claude>, anthropic API model claude-sonnet-4-6.
- [14] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: *International Conference on Machine Learning (ICML)*, 2017.
- [15] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: *International Conference on Machine Learning (ICML)*, 2005.
- [16] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [17] S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, S. Lyu, Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 4324–4333. doi:10.1109/CVPRW63382.2024.00436.
- [18] P. Yu, J. Fei, H. Gao, X. Feng, Z. Xia, C.-H. Chang, Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection, in: *International Conference on Machine Learning (ICML)*, 2025. ArXiv:2503.14853.
- [19] T. Li, Z. Huang, H. Wen, Y. He, S. Lyu, B. Wu, G. Cheng, RAIDX: A retrieval-augmented generation and GRPO reinforcement learning framework for explainable deepfake detection, in: *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, Dublin, Ireland, 2025. doi:10.1145/3746027.3754798, arXiv:2508.04524.
- [20] X. Guo, X. Song, Y. Zhang, X. Liu, X. Liu, Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Oral; arXiv:2503.20188.
- [21] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: *International Conference on Learning Representations (ICLR)*, 2015.
- [22] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: *International Conference on Learning Representations (ICLR)*, 2018.
- [23] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [24] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in:

- International Conference on Machine Learning (ICML), 2019.
- [25] P. Kawa, M. Plata, P. Syga, Defense against adversarial attacks on audio deepfake detection, in: Interspeech 2023, 2023, pp. 5276–5280. doi:10.21437/Interspeech.2023-409.
- [26] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, Y. Zhao, Forgery-aware adaptive transformer for generalizable synthetic image detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. ArXiv:2312.16649.
- [27] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, L. Yuan, DF40: Toward next-generation deepfake detection, in: Advances in Neural Information Processing Systems (NeurIPS), 2024. ArXiv:2406.13495.

A. Reproducibility Notes

Compute environment. Generation work performed on Paperspace NVIDIA RTX A6000 (47.5 GB VRAM); Detection work performed on Google Colab (CPU + T4 GPU as needed). All analysis notebooks runnable on CPU.

Hugging Face model IDs (verified).

- prithivMLmods/Deep-Fake-Detector-v2-Model (ViT)
- prithivMLmods/deepfake-detector-model-v1 (SigLIP)
- dima806/deepfake_vs_real_image_detection (ViT)
- umm-maybe/AI-image-detector (Swin)
- Organika/sdxl-detector (Swin)

Notebook order. NB-00 setup; NB-01 per-detector calibration; NB-02 orientation validation (with the cross-distribution table, Table 12); NB-03 $v1 \leftrightarrow v2$ comparison; NB-04 ablation studies (with all-normal ORIENTATIONS patch); NB-05 threshold sensitivity (same patch); NB-07 LaTeX table builder; NB-08 frequency forensics; NB-09 combined system-overview figure; NB-10 Wilson 95% binomial CIs on the official per-subset accuracies.

Calibration-set construction. 194 real frames sampled from the 335 dev videos; 100 InstantID fakes from our Submission 2; 100 affine-warp fakes from our Submission 1. Stored as `calibration_scores.csv` with columns `label`, `prithivMLmods_v2`, `prithivMLmods_siglip`, `dima806`, `umm_maybe`, `Organika`, `ash12321`, `proxy_eff_b0`, `freqms_b4`, `subtype`.

B. Heem2 Orientation: A Post-Submission Note

During reproducibility analysis (14 May 2026), we discovered that an internal artefact `audio_detector_features_v2-fix.csv` in our cache had inadvertently applied a $1 - x$ transform to Heem2’s output column. This file was *not* the source of our v2 audio submission. The submitted v2 audio uses the pre-fix variant `audio_detector_features.csv` with Heem2 in its original orientation. We document this here to preserve provenance: the submitted v2 audio is the simple average of MelodyMachine, Heem2, and Gustking in their as-released orientations, and the $v1 \leftrightarrow v2$ audio Pearson of +0.366 in Table 13 corresponds to that variant.