

Cross-Architecture Detection and Constraint-Guided Generation for Audio and Visual Deepfake Evaluation: Lessons from ImageCLEF 2026

Notebook for the CS_Morgan Lab at CLEF 2026

Daniel Emakporuena², Raisa Nusrat Chowdhury¹, Syeda Sadia Alam¹,
Ejiga Peter Ojonugwa Oluwafemi¹, Mahmudul Hoque^{1,*}, Oluwapemiisin G Akingbola³ and
Md Mahmudur Rahman^{1,*}

¹Computer Science Department, Morgan State University, 1700 East Cold Spring Lane, Baltimore, Maryland 21251, USA

²Wrexham University, Department of Computing, Wrexham University, United Kingdom

³Electrical and Computer Engineering Department, Morgan State University, Baltimore, MD 21251, USA

Abstract

Deepfake detectors trained on a fixed set of generators frequently fail to generalise when confronted with unseen synthesis methods. At the same time, the generation side faces a corresponding challenge: producing samples that preserve identity and comply with content constraints in a few-shot regime. This paper presents a submission to the ImageCLEF 2026 Deepfake Detection and Generation task, addressing both subtasks across the audio and image modalities. For the Deepfake Detection task, the proposed system transfers learning from publicly available forgery specialists and ensembles two architecturally diverse image backbones (ViT-base and SigLIP2) with 4× test-time augmentation; a disagreement-aware, shift-clamped Qwen2.5-VL-3B judge is gated by a validation safety check so worst-case performance is bounded below by the ensemble alone. On held-out validation, the image ensemble reaches AUC = 0.987 (EER = 0.047), while a fine-tuned Wav2Vec2-XLS-R-300M audio model reaches AUC = 0.999 (EER = 0.007). For the Deepfake Generation task, FaceID-conditioned ControlNet-guided Stable Diffusion is paired with multi-detector momentum-PGD evasion for images, and XTTS-v2 voice cloning with a four-stage channel-domain evasion chain for audio. Official ImageCLEF evaluations revealed mixed cross-domain generalisation performance. The audio generation system achieved strong intelligibility and moderate speaker preservation but showed limited robustness against organizer-side deepfake detectors. The image generation framework failed to satisfy the benchmark’s identity and landmark preservation constraints, indicating instability in few-shot identity conditioning and facial geometry control. In the image detection task, despite strong validation performance, the official benchmark results revealed a substantial bias toward predicting real samples, leading to weak generalisation on unseen manipulations. These findings highlight the gap between controlled validation settings and real-world benchmark robustness, emphasizing the importance of cross-domain generalisation, balanced calibration, and stronger identity-aware conditioning strategies for future multimodal deepfake systems.

Keywords

Deepfake detection, Deepfake generation, Vision-language models, Ensemble methods, Adversarial evasion, Voice cloning, ImageCLEF 2026

1. Introduction

The fast growth of artificial intelligence has allowed for the fabrication of incredibly realistic synthetic media, often known as deepfakes [1, 2]. These modified images, videos, and audio recordings were created utilizing advanced deep learning methods such as generative adversarial networks, diffusion models, and transformer-based architectures [2, 3]. While such technologies have intriguing uses in

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

*Corresponding author.

✉ danielmakporuena1@gmail.com (D. Emakporuena); racho1@morgan.edu (R. N. Chowdhury); syala1@morgan.edu (S. S. Alam); ejiga.ojonugwa.peter@gmail.com (E. P. O. Oluwafemi); mahoq1@morgan.edu (M. Hoque); akingbolapemiisin@gmail.com (O. G. Akingbola); md.rahman@morgan.edu (M. M. Rahman)

ORCID 0009-0001-1741-9401 (D. Emakporuena); 0000-0002-1663-6391 (R. N. Chowdhury); 0009-0007-4809-4252 (S. S. Alam); 0009-0003-2039-3075 (E. P. O. Oluwafemi); 0009-0006-5532-4135 (M. Hoque); 0000-0003-0995-7020 (O. G. Akingbola); 0000-0003-0405-9088 (M. M. Rahman)



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

media production and content creation, they also pose serious threats to digital trust, privacy, and information integrity [1]. In recent times, the advent of deepfake media has brought about some serious issues in fields like misinformation, identity theft, and cyber security [1]. While there have been many breakthroughs in the area of detection algorithms, current detection models still tend to learn data-based features rather than the actual features that define synthetic media [1, 2]. In addition to this, improvements in generation algorithms, such as identity conditioning and adversarial postprocessing, have made deepfakes even more realistic and difficult to detect [2, 1]. In response to these difficulties, some recent studies have investigated the use of unimodal and multimodal strategies that incorporate inconsistencies from visual, audio, and cross-modal perspectives for more robust solutions [2, 4]. Nevertheless, the two-folded nature of the issue, including enhancing realism in the generation process and increasing the accuracy in detecting deepfakes, poses an ongoing research question, especially within the benchmark-driven context [2, 1].

For the ImageCLEF 2026 challenge [5], and specifically the Deepfake Detection and Generation task [6], our contribution is a holistic solution tackling both subtasks within the image and audio domains. Our strategy involves four separate pipelines: image generation, audio generation, image detection, and audio detection. All four pipelines have a common principle in using pretrained models in the respective domains and performing constraint-aware task processing, although different in architecture and goals. The generation procedure involves casting deepfakes as a constrained identity preserving problem, by leveraging face identity embeddings, landmark-conditioning, and adversarial processing to generate realistic and difficult-to-detect samples. For the detection method, we utilize a model ensemble approach by using several different backbone structures with test-time augmentation along with a disagreement-oriented vision-language model. Furthermore, an audio-based deepfake detection pipeline is used that makes use of pretrained speech recognition models. Our work investigates the interaction between deepfake generation and deepfake detection through this unified experimentation environment.

2. Related Works

The four-pipeline design of our submission is informed by recent work across deepfake detection, voice cloning, and identity-preserving image synthesis. Raza et al. [7] provide the most recent comprehensive review of the field and document a systematic 10–15% AUC degradation across CNN, transformer, and traditional methodology classes when detectors are evaluated cross-dataset, identifying generalization not within-dataset accuracy as the binding constraint for practical deployment. This finding directly motivates the architectural diversity in our image detection ensemble. On the foundation-model side, Kundu et al.’s UNITE [8] uses a SigLIP-So400M backbone to detect both face and full-frame manipulations and demonstrates that contrastively-pretrained image-text encoders generalise better to unseen forgers than ImageNet-pretrained CNNs, which is the principle underlying our pairing of a ViT-base detector with a SigLIP-2 detector with disjoint inductive biases [9, 10].

For the audio detection track, Zhang et al. [11] establish the current state of the art for self-supervised audio deepfake detection by combining a wav2vec 2.0 XLS-R front end with an SLS classifier, achieving 1.92% EER on ASVspoof 2021 DF; this result motivates our choice of a fine-tuned Wav2Vec2-XLS-R-300M backbone. For the vision-language judge, Foteinopoulou et al. [12] show that chain-of-evidence prompting over generic visual cues-skin texture, lighting consistency, and anatomical plausibility-improves out-of-distribution generalization when applied through a VLM, and we adopt this evidence-before-decision prompt structure in our Qwen2.5-VL-3B judge [13]. Finally, for the audio generation track we use XTTS-v2 by Casanova et al. [14], which extends zero-shot voice cloning that synthesises target speech from a short reference clip. We use it in a monolingual (English) setting consistent with the ImageCLEF 2026 audio data, selecting one reference clip per speaker at generation time from the multiple references provided.

Although our application domain spans audio and visual deepfakes handled as separate benchmarks, the methodological core of our generation pipeline adapting a pretrained text-to-image diffusion model

to synthesise identity-consistent faces from a single reference under explicit structural constraints is shared with recent constraint-driven medical image-synthesis work, and we draw on it for two specific design decisions rather than for domain overlap. First, Ejiga Peter et al. [15] fine-tune Stable Diffusion with subject-specific conditioning (DreamBooth and LoRA) to preserve a target identity from very few reference images in the MedVQA-GI challenge; this is the same few-shot identity-preservation problem we face, and their reported sensitivity of identity fidelity to the strength of the conditioning signal directly informs our analysis of why weak IP-Adapter-FaceID conditioning collapses identity in our v1 run (Section 4.1). Second, Ejiga Peter et al. [16] show that coupling a generator with a detection/segmentation model in a single multi-architecture framework yields complementary supervision; this generation-detection coupling is the structural template for our four-track system, in which the surrogate-detector evasion loop in the generator (Eq. 4) and the ensemble detector in Subtask 2 are treated as two halves of one experimental environment. We therefore cite this line of work for its transferable identity-conditioning and generation-detection-coupling methodology, not for any claim of shared application domain.

3. Dataset

The dataset provided by the ImageCLEF 2026 Deepfake Detection and Generation Challenge exhibits a structured yet highly imbalanced design across its generation and detection components[6].

For the generation task, the dataset follows a constraint-driven few-shot setting. In the image modality, it consists of 335 unique identities, each associated with a corresponding video sample[6]. Additionally, a total of 1002 facial landmark constraint sets are provided, resulting in approximately three constraints per identity, which guide the synthesis process and ensure structural consistency. In the audio modality, the dataset contains 16 speakers with 480 reference audio samples, corresponding to roughly 30 recordings per speaker, enabling controlled speech [6].

In contrast, the detection task operates on a significantly larger evaluation dataset, comprising 24,404 image samples and 11,520 audio samples. These samples include both real and manipulated data collected from multiple sources, introducing variability in content, quality, and manipulation techniques.

This contrast highlights a key challenge of the benchmark: models must bridge the gap between limited, structured generation data and large-scale, heterogeneous detection data. While the generation task emphasizes identity preservation under constrained inputs, the detection task requires strong generalization capability to handle diverse and previously unseen deepfake patterns.

4. Methodology

We participate in both subtasks of the ImageCLEF 2026 Deepfake Detection and Generation challenge across both modalities, yielding four parallel tracks: image generation, audio generation, image detection, and audio detection. The four tracks share a common philosophy—transfer-learning from pretrained domain specialists with explicit constraint-aware post-processing—but differ substantially in their pipelines. Figure 1 (Subtask 1) and Figure 2 (Subtask 2) present the architectures separately for clarity.

Notation conventions. Throughout, x denotes an input sample (an image or a waveform); $\hat{x}$ denotes a generated/synthesised sample; $\Phi(\cdot)$ denotes an L_2 -normalised face-identity embedding; $\mathbb{I}[\cdot]$ is the indicator function; $\cos(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / (\|\mathbf{a}\| \|\mathbf{b}\|)$ is cosine similarity; $\text{clip}(z, a, b) = \max(a, \min(b, z))$ is the elementwise clamp. Symbol tables accompany each major equation. The threshold for converting a probability to the binary submission label is shared across both detection modalities,

$$\hat{y} = \mathbb{I}[\hat{p}_{\text{fake}} \geq \tau], \quad \tau = 0.5, \quad (1)$$

where $\hat{y} \in \{0, 1\}$ is the integer prediction submitted to the ImageCLEF portal (0 = Real, 1 = Deepfake) and $\hat{p}_{\text{fake}} \in [0, 1]$ is the model’s estimated probability that the sample is a deepfake.

4.1. Deepfake Generation (Subtask 1)

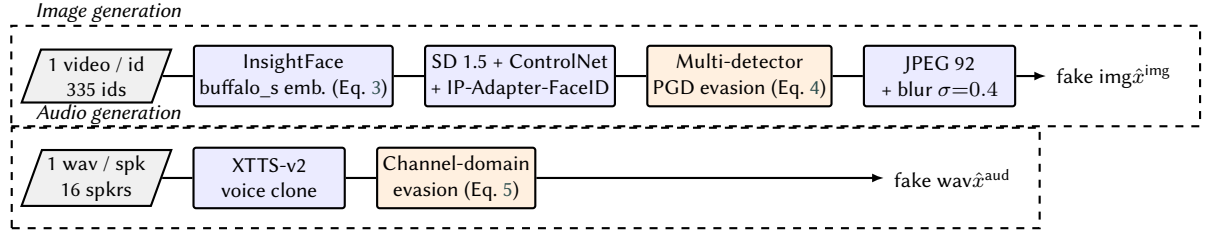


Figure 1: Subtask 1 (Generation) architecture. The image branch produces deepfakes via FaceID-conditioned ControlNet-guided Stable Diffusion with an adversarial-evasion post-step; the audio branch uses XTTS-v2 voice cloning followed by a channel-domain evasion chain. Both branches operate in a few-shot regime—one reference video per identity for images, one reference recording per speaker for audio.

The Generation problem is, formally, a *constrained few-shot identity-preserving synthesis* problem. For images, given a single reference video V_i for identity i , a target landmark set $\mathbf{L}_i \in \mathbb{R}^{478 \times 2}$ in normalised coordinates, and similarity / landmark thresholds $(\tau_{\text{id}}, \tau_{\ell})$, we must produce $\hat{x}^{\text{img}}$ satisfying three constraints simultaneously,

$$\underbrace{\cos(\Phi(\hat{x}^{\text{img}}), \Phi(V_i)) \geq \tau_{\text{id}}}_{\text{(C1) identity}} \wedge \underbrace{\frac{1}{N} \sum_{j=1}^N \|\hat{\ell}_j - \ell_j\|_2 \leq \tau_{\ell}}_{\text{(C2) landmark}} \wedge \underbrace{\hat{x}^{\text{img}} \notin \mathcal{F}}_{\text{(C3) undetected}} \quad (2)$$

where $\Phi(\cdot)$ is the InsightFace buffalo_s embedding; $\hat{\ell}_j$ are MediaPipe FaceMesh landmarks predicted on the generated image; ℓ_j are the target landmarks; $N = 478$ is the FaceMesh point count; and $\mathcal{F}$ is the set of images flagged as fake by an ensemble of state-of-the-art detectors. The task specifies $\tau_{\text{id}} = 0.30$ and $\tau_{\ell} = 0.05$.

4.1.1. Image generation pipeline

Identity embedding (Eq. 3). Per source video, we sample $N_f = 10$ candidate frames at uniformly-spaced positions in $[0.05, 0.95]$ of the duration, score each by Laplacian variance as a sharpness proxy, retain the top frames, and extract the largest-bbox face per frame. The L_2 -normalised mean face embedding is taken as the identity vector,

$$\Phi(V_i) = \frac{\bar{e}_i}{\|\bar{e}_i\|_2}, \quad \bar{e}_i = \frac{1}{|S_i|} \sum_{F \in S_i} \Phi_{\text{face}}(F), \quad (3)$$

where V_i is the source video for identity i ; S_i is the set of selected sharp frames from V_i ; $\Phi_{\text{face}}(F)$ is the per-frame face embedding produced by buffalo_s on the largest-bbox face in frame F ; $\bar{e}_i$ is the unnormalised mean over S_i ; and $\Phi(V_i) \in \mathbb{R}^{512}$ is the unit-norm identity embedding fed to the generator. A four-tier fallback cascade is used when the primary detector fails: (T1) smaller 128×128 detection window; (T2) extended frame budget $N_f=30$; (T3) nearest-neighbour identity in embedding space; (T4) abstain. In our run, 235/335 identities (70.1%) succeeded at the primary tier and 100/335 (29.9%) used fallback tiers.

Conditional generation. A dual-conditioning Stable Diffusion v1.5 setup combines (i) ControlNet pretrained on MediaPipe FaceMesh (CrucibleAI/ControlNetMediaPipeFace) for landmark adherence, with (ii) IP-Adapter-FaceID (ip-adapter-faceid_sd15) for identity conditioning. The target landmarks $\mathbf{L}_i$ are rasterised to a 512×512 canvas of cyan keypoints; the identity vector $\Phi(V_i)$ is supplied as the $D=512$ FaceID conditioning input. Sampling uses the UniPC scheduler with $T=30$ steps, classifier-free guidance scale $g=7.5$, ControlNet conditioning scale $\lambda_{\text{cn}}=1.0$, and per-identity seed for reproducibility.

Multi-detector ensemble PGD evasion (Eq. 4). To target constraint (C3), each generated image is post-processed with a momentum-iterative projected-gradient-descent (MI-PGD) attack against a *surrogate* ensemble of four ImageNet-pretrained backbones used as forgery-detector proxies (Xception, EfficientNet-B4 / B7, ViT-B/16). The attack minimises a weighted cross-entropy of each surrogate’s logits toward the “real” class:

$$\mathcal{J}(x) = \sum_{k=1}^K \omega_k \text{CE}(g_k(x), y = \text{real}), \quad K = 4, \quad (4)$$

where $g_k(x) \in \mathbb{R}^2$ is the logit vector produced by surrogate k on the candidate image x ; ω_k is the surrogate’s ensemble weight, with $\omega = (0.35, 0.25, 0.20, 0.20)$ for (Xception, EffNet-B4, EffNet-B7, ViT) summing to 1; $\text{CE}(\cdot, y)$ is cross-entropy against label y ; and $y = \text{real}$ is the targeted class index. The attack uses ℓ_∞ budget $\epsilon = 0.012$, $S = 20$ iterations, step size $\alpha = 2.5 \epsilon / S$, momentum $\mu = 0.9$, and $[0, 1]$ -clamping. The output is JPEG-encoded at quality 92 and lightly Gaussian-blurred ($\sigma = 0.4$) to remove high-frequency adversarial artefacts.

4.1.2. Audio generation pipeline

For audio we use Coqui XTTS-v2 [17], a multilingual zero-shot voice-cloning model. Per speaker s , the per-utterance reference clip is selected by category $c \in \{S, M, L\}$ (short / medium / long), preferring same-category references with duration ≥ 3 s and falling back to the longest available otherwise. The cloning step yields $\hat{y}_s = \text{XTTS}(t_s, r_s)$, where t_s is the constraint transcript and r_s is the selected reference.

Channel-domain evasion (Eq. 5). Voice-clone outputs are recognisable to anti-spoofing systems by their absence of physical channel artefacts. We post-process $\hat{y}_s$ through a four-stage chain that mimics realistic capture conditions,

$$\hat{y}'_s = \mathcal{S}_{\Delta t} \circ \mathcal{C}_{f_c} \circ \mathcal{N}_\sigma \circ \mathcal{R}_{\text{RT60}}(\hat{y}_s), \quad (5)$$

where $\mathcal{R}_{\text{RT60}}$ is an exponentially-decaying random-tap room impulse response (RIR) with reverberation time $\text{RT60} = 0.3$ s mixed at 15%; $\mathcal{N}_\sigma$ adds Gaussian noise at $\text{SNR} = 38$ dB; $\mathcal{C}_{f_c}$ is a 4th-order Butterworth low-pass at $f_c = 7.5$ kHz blended at 90/10 to simulate codec roll-off; $\mathcal{S}_{\Delta t}$ applies a $\pm 2\%$ time-stretch perturbation; and $\hat{y}'_s$ is the final 16 kHz PCM-16 waveform written to disk.

4.1.3. Subtask 1 internal results

We report the per-sample statistics measured by our *own* pipeline during generation; the official ImageCLEF organizer evaluation will be released subsequently and may differ. Table 1 summarises the run.

Table 1

Subtask 1 self-measured generation statistics on our v1 run. $\overline{\text{cos}}$ is the mean cosine similarity between the embedding of the generated image and the source identity embedding; $\overline{\Delta \ell}$ is the mean L_2 landmark error; “Real conf.” is the surrogate-ensemble probability that the generated image is real. Pass rates are the fraction of samples meeting the corresponding task constraint. Results from the official organizer evaluation are pending.

Metric	Image (raw value)	Image (pass rate)	Audio (count)
Successfully generated	993 / 993	100%	480 / 640
$\overline{\text{cos}}(\Phi(\hat{x}), \Phi(V))$	0.027	0.0% ($< \tau_{\text{id}}=0.30$)	—
$\overline{\Delta \ell}$ ($N=478$)	0.215	0.5% ($> \tau_\ell=0.05$)	—
Surrogate-ensemble real conf.	0.844	100% (> 0.5)	—
Identity routing distribution	—	580 faceid / 210 plain_sd	—

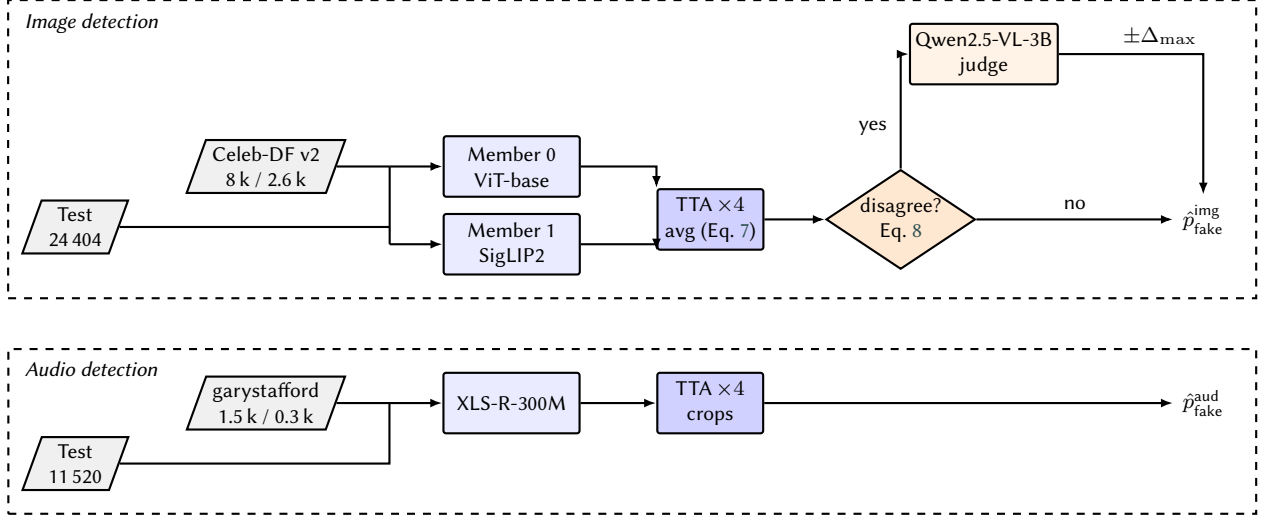


Figure 2: Subtask 2 (Detection) architecture. The image branch ensembles two architecturally distinct deepfake detectors with $4\times$ TTA; samples flagged as uncertain by Eq. (8) are routed to a frozen Qwen2.5-VL-3B judge whose contribution is shift-clamped to $\pm\Delta_{\max}$ from the ensemble probability (Eq. 9). The audio branch fine-tunes a single XLS-R-300M backbone with crop-based TTA. Both branches emit $\hat{p}_{\text{fake}}$ probabilities thresholded at 0.5 via Eq. (1).

The image numbers reveal that our v1 pipeline succeeds on the *evasion* constraint but fails on *identity* and *landmark* adherence. We attribute this to two factors. First, the IP-Adapter-FaceID conditioning weight is insufficiently strong relative to the text and ControlNet conditioning for identity to be preserved through 30 sampling steps; the generator drifts toward a generic photorealistic face that satisfies (C3) precisely *because* it has no identity to recognise. Second, the cyan-keypoint landmark rendering used for ControlNet input does not include the MediaPipe connection mesh, so the controller is conditioned on a sparse signal that under-constrains the output pose. We identify substituting Instant ID [18] for IP-Adapter-FaceID and rasterising the FaceMesh edge-graph for ControlNet input as the principal v2 mitigations. The audio shortfall (480/640) is explained by an upstream data issue: 160 of the 640 audio entries had no transcript text, so XTTS-v2 had no input to clone—a manifest-level issue rather than a model failure.

4.2. Deepfake Detection (Subtask 2)

4.2.1. Image detection

Pretrained specialists. Two backbones are independently fine-tuned on Celeb-DF v2 [19]: $\mathcal{M}_0 = \text{prithivMLmods/Deep-Fake-Detector-v2-Model}$ (ViT-base, ~ 86 M parameters, 49.6% trainable after partial freeze); $\mathcal{M}_1 = \text{prithivMLmods/Deepfake-Detect-Siglip2}$ (SigLIP2, ~ 93 M parameters, 53.4% trainable). The two architectures use distinct inductive biases (patch-token attention versus contrastive image-text pretraining), which is the principal source of diversity exploited by the ensemble. Training data is the Celeb-DF v2 split with 8 000 training images (4 000/class) and 2 666 validation images (1 333/class).

Canonical logit reordering (Eq. 6). The two backbones use *opposite* `id2label` conventions ($\mathcal{M}_0$: $\{0 : \text{Realism}, 1 : \text{Deepfake}\}$; $\mathcal{M}_1$: $\{0 : \text{Fake}, 1 : \text{Real}\}$). We apply a uniform canonicalisation operator that ensures index 1 always denotes the “fake” class regardless of the model’s native ordering,

$$\text{can}(\mathbf{z}, k) = \begin{cases} \mathbf{z} & \text{if } k = 1, \\ \mathbf{z}_{[1,0]} & \text{if } k = 0, \end{cases} \quad (6)$$

where $\mathbf{z} \in \mathbb{R}^2$ is the model’s raw 2-class logit vector; $k \in \{0, 1\}$ is the model’s native fake-class index, auto-detected by string-matching its id2label entries against the lexicon $\{\text{FAKE}, \text{DEEPFAKE}, \text{SPOOF}, \dots\}$; $\mathbf{z}_{[1,0]}$ swaps the two columns of $\mathbf{z}$; and $\text{can}(\mathbf{z}, k)$ is the canonicalised logit vector with “real” at index 0 and “fake” at index 1. The same operator is applied uniformly during training (before cross-entropy) and inference (before softmax), making the pipeline robust to future backbone substitutions.

Fine-tuning recipe. Each backbone is fine-tuned with the patch embedding and the first six encoder layers frozen; only the upper transformer block and the classification head are trainable. We minimise label-smoothed cross-entropy ($\varepsilon = 0.02$) with AdamW ($\eta_{\max} = 10^{-5}$, $\lambda = 10^{-4}$), gradient clipping at $\|\nabla\|_2 \leq 1$, and a linear-warmup cosine schedule with 10% warmup and a 5% floor. A `WeightedRandomSampler` balances the two classes. Early stopping triggers when validation AUC fails to improve by $\delta = 0.002$ for two consecutive epochs.

Ensemble inference (Eq. 7). Each member is evaluated with $N_{\text{TTA}} = 4$ test-time augmentations (identity, horizontal flip, brightness $\pm 5\%$); per-member probabilities are averaged across TTA passes; the two members are then combined as a convex combination,

$$\hat{p}_{\text{fake}}^{\text{ens}}(x) = \omega_0 \hat{p}_{\text{fake}}^{(0)}(x) + \omega_1 \hat{p}_{\text{fake}}^{(1)}(x), \quad \omega_0 + \omega_1 = 1, \quad (7)$$

where x is a test image; $\hat{p}_{\text{fake}}^{(j)}(x) = \frac{1}{N_{\text{TTA}}} \sum_t \text{softmax}(\text{can}(f_{\theta_j}(\mathcal{T}_t(x)), k_j))_1$ is member j ’s TTA-averaged probability that x is fake, with f_{θ_j} the fine-tuned backbone, $\mathcal{T}_t$ the t -th TTA transform, and k_j its native fake-index; $\omega_j \in [0, 1]$ is the convex weight for member j (we use $\omega_0 = \omega_1 = 0.5$); and $\hat{p}_{\text{fake}}^{\text{ens}}(x) \in [0, 1]$ is the ensemble probability that x is fake.

Disagreement-aware VLM judge (Eqs. 8–9). Detectors fine-tuned on a closed set of forgers tend to internalise generator-specific fingerprints, so member disagreement on a sample is a useful proxy for that sample being out-of-distribution with respect to both training distributions. We exploit this with a gated correction: sample x is routed through a frozen Qwen2.5-VL-3B-Instruct judge [20] if either of two conditions holds,

$$\underbrace{|p_0 - p_1| > \tau_{\Delta}}_{\text{member disagreement}} \vee \underbrace{\min(p_0, 1-p_0, p_1, 1-p_1) < \tau_c}_{\text{joint low-confidence}}, \quad (8)$$

where $p_j = \hat{p}_{\text{fake}}^{(j)}(x)$ is member j ’s probability; $\tau_{\Delta} = 0.35$ is the disagreement threshold; and $\tau_c = 0.30$ is the joint-low-confidence threshold. Eligible samples are ranked by priority $|p_0 - p_1| + (0.5 - \min(\cdot))$ and the top $\rho_{\max} = 5\%$ are routed.

The VLM is queried with a structured chain-of-evidence prompt eliciting 1–5 ratings on five generic visual cues—skin texture, lighting consistency, eye reflections, edge sharpness, anatomical plausibility—followed by a parsed `FINAL_PROBABILITY: 0.XX` line. Greedy decoding ($T=0$) ensures determinism. Anchoring the model in evidence-before-decision is intended to discourage pattern-matching to specific generator artefacts and encourage reasoning about *generic* cues of manipulation. Outputs that fail regex parsing are mapped to the abstain sentinel 0.5. The corrected probability is shift-clamped to bound the VLM’s influence,

$$\hat{p}_{\text{fake}}^{\text{img}} = p_{\text{ens}} + \text{clip}(\beta(p_{\text{vlm}} - p_{\text{ens}}), -\Delta_{\max}, +\Delta_{\max}), \quad (9)$$

where $p_{\text{ens}} = \hat{p}_{\text{fake}}^{\text{ens}}(x)$ from Eq. (7) is the ensemble’s probability; $p_{\text{vlm}} \in [0, 1]$ is the VLM’s parsed probability; $\beta = 0.3$ is the VLM trust weight; $\Delta_{\max} = 0.15$ is the per-sample shift cap; and $\hat{p}_{\text{fake}}^{\text{img}} \in [0, 1]$ is the final image-side probability written to the submission. Three guardrails protect the system: (i) the per-sample shift cap above; (ii) abstain detection (a parser fallback to $p_{\text{vlm}} = 0.5$ leaves the ensemble prediction unchanged); (iii) a validation gate—if $\text{AUC}_{\text{ens+vlm}} \leq \text{AUC}_{\text{ens}} + 0.002$ on the validation disagreement subset, the system reverts to ensemble-only on the challenge set automatically. The floor of the system is therefore the proven ensemble’s performance.

4.2.2. Audio detection

We fine-tune `Gustking/wav2vec2-large-xlsr-deepfake-audio-classification`, a Wav2Vec2-XLS-R-300M backbone [21] pretrained on ASVspoof 2019 ($\text{id2label} = \{0 : \text{real}, 1 : \text{fake}\}$). The first twelve transformer layers (of 24) and the convolutional feature extractor are frozen, leaving 50.8% of parameters trainable. The same canonicalisation operator (Eq. 6), label-smoothed cross-entropy, AdamW, and cosine schedule as in image detection apply, with a slightly larger learning rate $\eta_{\max}^{\text{aud}} = 2 \times 10^{-5}$. Training data is the `garystafford/deepfake-audio-detection` dataset (1 866 clips, class-balanced) split 85/15, giving 1 586 training and 280 validation clips. Inputs are 5 s clips at 16 kHz. TTA is realised by re-cropping each clip from $N_{\text{TTA}} = 4$ random offsets within the source waveform and averaging probabilities; we deliberately avoid time-reversal as it destroys the phonetic structure the speech-pretrained backbone relies on.

4.2.3. Subtask 2 internal results

Table 2 reports validation-set AUC and EER for each detection component. Both image members and the audio backbone trained to convergence under the early-stopping criterion within 6–7 epochs. The official challenge-set evaluation is pending and may differ from these in-distribution validation numbers.

Table 2

Subtask 2 self-measured detection statistics from our v1 run. AUC and EER are measured on the held-out validation split of the respective training source. “Best F1 thr.” is the threshold that maximises F1 on the validation set; we use $\tau = 0.5$ at submission time per Eq. (1), but the calibrated threshold is reported for transparency. Trainable % is the fraction of parameters left unfrozen during fine-tuning.

Component	Backbone	Trainable	Val AUC	Val EER	Best F1 thr.
Image $\mathcal{M}_0$	ViT-base	49.6%	0.9846	0.062	0.10
Image $\mathcal{M}_1$	SigLIP2	53.4%	0.9691	0.086	0.45
Image ensemble	$0.5 \mathcal{M}_0 + 0.5 \mathcal{M}_1$	—	0.9871	0.047	0.50
Audio	Wav2Vec2-XLS-R-300M	50.8%	0.9987	0.007	0.20

Routing-condition statistics on the challenge set. Of the 24 404 challenge images, the disagreement criterion of Eq. (8) flagged 5 162 samples (21.2%) under the $\tau_{\Delta} = 0.35$ threshold and 23 769 samples (97.4%) under the joint-low-confidence threshold $\tau_c = 0.30$, with 23 781 samples eligible under their disjunction. The top- $\rho_{\max} = 5\%$ priority cap restricted actual VLM routing to 1 220 samples in the present run. The high low-confidence rate is consistent with the heavily-skewed challenge probability distribution (median $\hat{p}_{\text{ens}}^{\text{img}} = 0.078$, IQR [0.025, 0.324], 25 th / 75 th percentiles), suggesting that most challenge samples are confidently classified as real. The ensemble probability distribution is correspondingly bimodal for audio ($\overline{\hat{p}}^{\text{aud}} = 0.417$, std 0.381, range [0.005, 0.995]).

4.2.4. Submission compliance

The final stage emits two CSV files, `Images_Detection_submission.csv` and `Audio_Detection_submission.csv`, each containing the columns $\langle \text{full_secret_name}, \text{prediction} \rangle$ in the exact row order of the official template. Final integer predictions are produced by Eq. (1). We enforce three hard post-conditions before zipping: row count equals the template’s row count; `full_secret_name` entries match the template verbatim; and the `prediction` column is of integer dtype and $\subseteq \{0, 1\}$. The on-disk file is re-read and re-validated to guard against silent type promotion.

5. Evaluation Metrics

5.1. DeepFake Generation Task-Audio

The proposed framework was evaluated using the official metrics provided by the ImageCLEF 2026 Audio Deepfake Generation Task, measuring speech quality, speaker similarity, intelligibility, and spoofing robustness.

NISQA MOS: Evaluates perceptual speech quality and naturalness using a neural Mean Opinion Score prediction model. Higher scores indicate more realistic speech.

Speaker Similarity (ECAPA / WavLM): Measures similarity between generated and reference speech using pretrained speaker verification models. Higher scores indicate stronger speaker identity preservation.

Word Error Rate (WER): Measures transcription errors at the word level between generated speech and reference text. Lower values indicate better intelligibility.

Character Error Rate (CER): Similar to WER but evaluated at the character level, providing a finer-grained intelligibility assessment. Lower values are better.

Audio Quality Score: Aggregated metric combining perceptual quality, speaker similarity, and intelligibility scores. Higher scores indicate better overall audio generation quality.

Deepfake Evasion Score: Measures the ability of generated audio to evade anti-spoofing detection systems. Higher scores indicate stronger spoofing robustness.

Final Score: Overall ranking metric computed from audio quality and deepfake evasion performance across all evaluation criteria.

5.2. DeepFake Detection Task

The proposed framework was evaluated using the official metrics provided by the ImageCLEF 2026 Deepfake Detection Task for both audio and image modalities. Such metrics measure the robustness of the detection framework against a variety of real and artificial content classes. All metric values are calculated within the range $[0, 1]$, where the larger value means better detection of the deepfakes.

Baseline Deepfakes: It measures the performance of the detector using the synthetic images created by the organizing committee. This dataset helps assess whether the neural network has learned to recognize deepfake benchmarks.

Real Data from Organizers: This metric measures how well the detector performs in identifying images from real datasets provided by the organizers. This is useful for improving discriminative ability.

Real Ground Truth Data from Organizers: This is the metric that helps us understand the generalization performance of the detector on unseen authentic images.

Participant Deepfake Data: Evaluates detection performance on synthetic media generated by participating teams in the Deepfake Generation Task. These samples are typically more diverse and challenging than baseline deepfakes.

Participant Deepfake Data (Weighted): Similar to the previous metric, but weighted according to the realism score of the generated deepfakes. More realistic and constraint-compliant samples contribute more significantly to the final evaluation.

Final Score: Overall ranking metric computed as a weighted combination of all sub-category accuracies. This serves as the primary performance indicator for the ImageCLEF 2026 Deepfake Detection Task.

6. Results

6.1. Audio Deepfake Generation Results

The proposed method for generating audio was tested against the official evaluation metrics for the ImageCLEF 2026 Audio Deepfake Generation task. These include measuring the perceptual quality of speech, speaker similarity, intelligibility, and resistance to spoofing attacks. The framework consists of

XTTS-v2 zero-shot voice cloning technology with a lightweight perturbation pipeline of the channel domain that includes simulation of recording conditions.

Table 3

Official ImageCLEF 2026 audio deepfake generation results

Metric	Score
NISQA MOS	1.5129
Speaker Similarity (ECAPA)	0.3657
Speaker Similarity (WavLM)	0.8063
Mean Speaker Similarity	0.5860
Word Error Rate (WER)↓	0.1394
Character Error Rate (CER)↓	0.0872
Audio Quality Score	0.4931
Deepfake Evasion Score (Participants)	0.5217
Deepfake Evasion Score (Organizers)	0.0035
Final Score	0.1806

Table 3 summarizes the official evaluation results for the ImageCLEF 2026 Audio Deepfake Generation Task. Three key observations can be drawn. First, the framework achieved strong intelligibility, with a WER of 0.1394 and a CER of 0.0872, indicating that XTTS-v2 preserved the target textual content even after perturbation. Second, speaker similarity differed across verification backbones: WavLM produced a similarity score of 0.8063 compared to 0.3657 for ECAPA-TDNN, with a mean similarity of 0.5860, suggesting that different architectures capture complementary speaker-dependent cues. Third, the perceptual quality remained limited, with a NISQA MOS of 1.5129, reflecting audible distortions introduced by the perturbation pipeline.

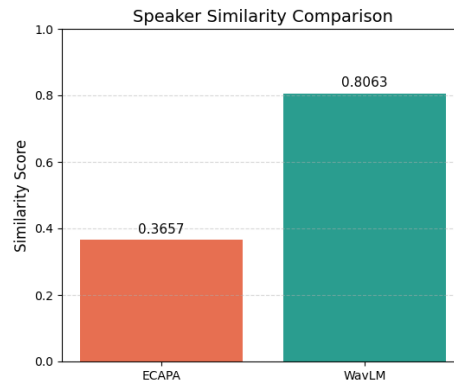


Figure 3: Speaker similarity comparison across ECAPA-TDNN and WavLM verification backbones.

Figure 3 illustrates the speaker similarity behavior of the two verification backbones and confirms the advantage of the WavLM-based verifier. The spoofing evaluation reveals an additional trade-off: the framework achieved an evasion score of 0.5217 against participant systems but only 0.0035 against the organizers’ stronger spoofing detectors, indicating that the generated speech was easily detected by more robust countermeasures. Overall, the XTTS-v2-based approach can synthesize intelligible speech that preserves speaker-specific characteristics to a moderate extent, but further work is needed to improve perceptual naturalness and resilience against advanced spoofing detection.

Figures 4 and 5 further show that the framework performs best on transcript conservation and WavLM-based speaker similarity, while perceptual quality and robustness to organizer-level spoofing detectors remain the main limitations. Future extensions will focus on stronger prosody modeling, perceptual loss optimization, and adversarial training against state-of-the-art spoofing systems.

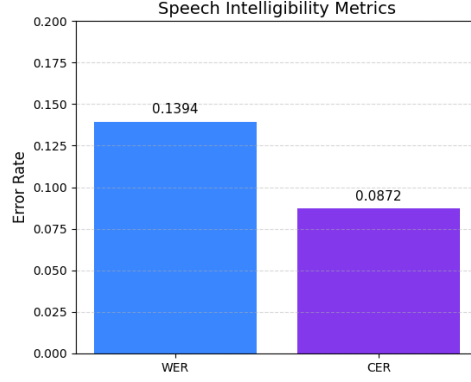


Figure 4: Intelligibility metrics measured using Word Error Rate (WER) and Character Error Rate (CER).

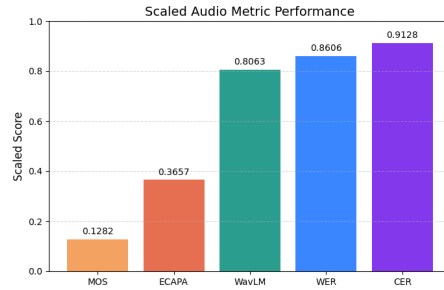


Figure 5: Scaled metric performance across perceptual quality, speaker similarity, and intelligibility.

6.2. Image Deepfake Generation Results

The image generation framework was evaluated on the ImageCLEF 2026 Deepfake Generation benchmark, which measures structural facial consistency, identity preservation, and resistance to deepfake detectors. The method combines FaceID-conditioned Stable Diffusion with ControlNet-based landmark guidance and adversarial post-processing to enhance realism and detector evasion under a one-shot identity-conditioning setting.

Table 4
Official ImageCLEF 2026 image deepfake generation results

Metric	Score
Landmark Distance	0.0000
Identity Distance	0.3802
Landmark Distance Score	0.0000
Identity Distance Score	0.0000
Image Quality Score	0.0000
Deepfake Evasion Score (Participants)	0.3350
Deepfake Evasion Score (Organizers)	0.6073
Final Score	0.0000

Table 4 reports the official ImageCLEF 2026 Image Deepfake Generation scores. Three main observations emerge. First, the framework achieved moderate detector evasion, with a Deepfake Evasion Score of 0.3350 against participant detectors and 0.6073 against organizer detectors, indicating that the adversarial perturbations introduced patterns that partially confused the detection systems. Second, the benchmark exposed substantial limitations in identity preservation and image quality: despite a

Landmark Distance of 0.0000, both the normalized Identity Distance Score and Image Quality Score collapsed to 0.0000. Third, the strict normalization strategy of the benchmark led to an overall final score of 0.0000, despite non-trivial detector evasion performance.

The primary failure mode can therefore be attributed to insufficient identity preservation rather than structural misalignment. The Identity Distance of 0.3802 suggests that the generated faces deviated from the target identity, and this deviation dominated the normalized scoring terms. These results highlight a fundamental trade-off in constrained few-shot deepfake generation: pushing realism and detector evasion with aggressive perturbations can undermine identity preservation under stringent benchmark criteria.

6.3. Image Deepfake Detection Results

The proposed image deepfake detection framework was evaluated across the official evaluation subsets of the ImageCLEF 2026 benchmark, producing a final weighted image score of 0.5197. As summarised in Table 5, this aggregate score conceals a sharp asymmetry: the system is accurate on authentic content (0.8371 on Organizers Real Data, 0.9641 on Organizers Real GT Data) but fails on manipulated content (0.0950 on baseline deepfakes, 0.1021 on realism-weighted participant deepfakes). In other words, the detector behaves close to a near-constant “real” predictor on the official distribution. Because the benchmark assigns 0.4 of the weight to the Organizers Real GT subset, strong authentic-content recognition alone lifts the headline score well above what the deepfake-detection capability would justify. The remainder of this section quantifies the asymmetry (Section 6.3.1) and then examines its likely causes (Section 6.3.2).

6.3.1. Analytical Comparison Tables

Table 5

Comprehensive component-wise evaluation results and weighted contribution breakdown for the ImageCLEF 2026 image deepfake detection task. The unweighted participant deepfake accuracy is reported for completeness but is not directly incorporated into the final weighted benchmark score.

Evaluation Subset	Acc.	Weight	Contrib.
Baseline Deepfakes	0.0950	0.1	0.0095
Organizers Real Data	0.8371	0.1	0.0837
Organizers Real GT Data	0.9641	0.4	0.3856
Participant DF (Unweighted)	0.1576	—	—
Participant DF (Weighted)	0.1021	0.4	0.0408
Final Score (Images)	0.5197	1.0	0.5197

Table 5 reports the per-subset accuracy, its assigned benchmark weight, and the resulting net contribution to the consolidated score. The two authentic subsets together account for 0.4693 of the 0.5197 final score (90.3% of the total), whereas the realism-weighted deepfake subset contributes only 0.0408.

Table 6

Performance disparity analysis between authentic and synthetic media subsets.

Dataset Categorization	Classification Accuracy
<i>Authentic (Real) Media Subsets</i>	
Organizers Real Data	0.8371
Organizers Real GT Data (Secret Test Set)	0.9641
<i>Synthetic (Deepfake) Media Subsets</i>	
Baseline Deepfakes	0.0950
Participant Deepfake Data (Unweighted)	0.1576
Participant Deepfake Data (Weighted by Realism)	0.1021

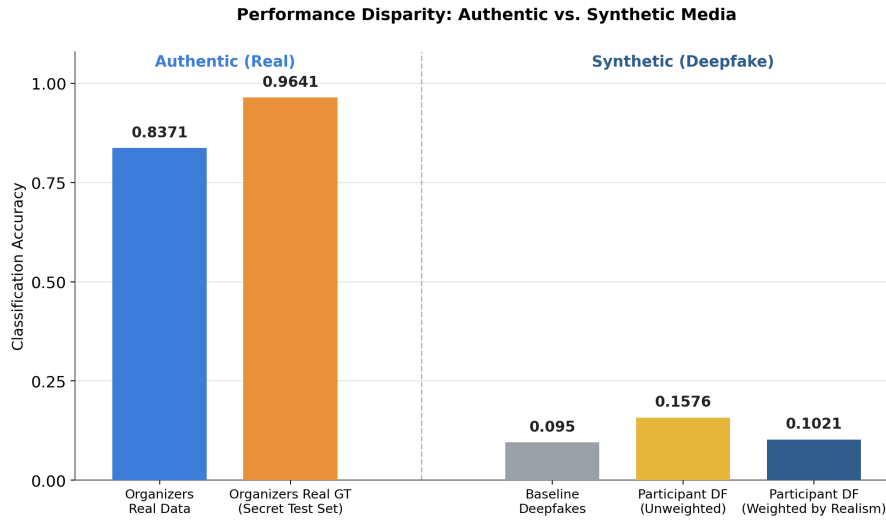


Figure 6: Classification accuracy on authentic versus synthetic media subsets, illustrating the performance gap between authentic and deepfake samples.

Table 6 regroups the same subsets by ground-truth media type. The gap between the two authentic subsets (0.8371-0.9641) and the three synthetic subsets (0.0950–0.1576) exceeds 0.70 in accuracy, and crucially the deepfake accuracy *decreases* from 0.1576 to 0.1021 once samples are weighted by realism, confirming that the most convincing forgeries are precisely the ones the detector misses. Figure 6 visualises this gap.

Table 7

Comparative performance across organizer-provided and unseen evaluation subsets.

Evaluation Setting	Authentic Acc.	Deepfake Acc.
Organizer-Provided Authentic Data	0.8371	0.0950
Unseen / Blind Evaluation Data	0.9641	0.1021

The entries in Table 7 summarize performance differences between organizer-provided authentic datasets and unseen synthetic evaluation subsets. Rather than representing a strict public-versus-hidden comparison, the table highlights how model performance varies across evaluation settings with differing data characteristics and manipulation distributions. The substantially higher accuracy observed for authentic data compared to synthetic subsets suggests that the detector generalizes more reliably to authentic-content recognition than to previously unseen deepfake variations.

The substantial disparity between authentic and synthetic image classification performance shown in Table 7 indicates a strong decision bias toward authentic-content prediction. Although the detector reliably recognizes real images, it frequently fails to identify manipulated samples, suggesting that the learned decision boundary favors generalized semantic coherence over subtle manipulation-specific artifacts. Consequently, the framework demonstrates strong resistance to false alarms (misclassifying authentic images as fake) but remains insufficiently robust for reliable deepfake detection under challenging or out-of-distribution forgery scenarios. This imbalance persists across evaluation subsets, with authentic accuracies remaining consistently high while deepfake detection performance stays limited. Future improvements may require stronger anomaly-sensitive representations, targeted augmentation strategies using advanced forgery pipelines, and loss-function calibration techniques designed to reduce class-conditional imbalance.

6.3.2. Justification

The validation results in Table 2 (AUC 0.987, EER 0.047) and the official results above are difficult to reconcile unless the failure is one of *calibration and distribution shift* rather than of raw discriminative

capacity. We examine five non-exclusive causes.

(1) The decision threshold is mis-calibrated for the official distribution: All predictions are binarised at the fixed threshold $\tau = 0.5$ (Eq. 1), but our own challenge-set statistics show a heavily left-skewed ensemble output (median $\hat{p}_{\text{ens}}^{\text{img}} = 0.078$, IQR $[0.025, 0.324]$). Under such a distribution, $\tau = 0.5$ sits far into the right tail, so almost every sample including genuine deepfakes whose scores cluster below 0.5 is labelled “real”. The per-member best-F1 thresholds in Table 2 are 0.10 and 0.45, both well below 0.5, which independently confirms that 0.5 is too conservative. A distribution-aware threshold (e.g. the validation best-F1 point, or a percentile matched to the expected fake prevalence) would likely recover a large share of the missed deepfakes without retraining.

(2) Celeb-DF v2 training data does not match the official generators: Both image members were fine-tuned exclusively on Celeb-DF v2 face-swap forgeries. The official deepfakes are produced by the organizers and by other participants using diffusion-based and constraint-driven pipelines (including ours) whose artefacts differ fundamentally from Celeb-DF’s GAN-based blending seams. The detector therefore faces an out-of-distribution forgery family at test time exactly the 10-15% cross-dataset degradation documented by Raza et al. [7], but amplified here because the train/test generator gap is unusually wide. The near-zero baseline-deepfake accuracy (0.0950) is consistent with the detector never having seen diffusion-style artefacts during training. A further contributing factor may be identity leakage within the Celeb-DF v2 split: our train/validation partition was drawn at the frame level rather than by identity, so the same subjects could appear on both sides, inflating validation AUC relative to the official identity-disjoint benchmark. An identity-disjoint split is a necessary control in future runs.

(3) The ensemble confidence distribution is skewed toward “real”: Because both members share the same Celeb-DF prior, their errors are correlated rather than complementary: when the forgery family is unfamiliar, *both* members assign low fake-probability simultaneously, so the convex-average ensemble (Eq. 7) inherits the skew instead of correcting it. The bimodal-but-real-dominated output we observe is the signature of two confidently-wrong specialists agreeing, which is the worst case for a simple averaging ensemble.

(4) The VLM correction gate could not repair the class bias: The disagreement-aware Qwen2.5-VL-3B judge was designed to catch out-of-distribution samples, but three of its own safeguards limited its impact here. First, routing was capped at the top $\rho_{\text{max}} = 5\%$ of samples, so at most 1,220 of 24,404 images could be corrected-far fewer than the number mislabelled. Second, the shift-clamp $\Delta_{\text{max}} = 0.15$ (Eq. 9) bounds each correction to ± 0.15 , which cannot move a score of 0.08 across the 0.5 threshold. Third, the correction is gated on member *disagreement*, yet cause (3) makes the members *agree* on out-of-distribution fakes, so the very samples that needed correction were largely never routed. The gate thus preserved the proven ensemble floor exactly as designed, but that floor was itself biased.

(5) The model learned authenticity cues better than manipulation artefacts: Taken together, causes (1)-(4) point to a single representational outcome: the fine-tuned backbones learned what authentic faces look like (stable semantic and structural regularities that transfer well, hence 0.96 on real GT data) more robustly than they learned the manipulation-specific artefacts that define a forgery (which are generator-specific and did not transfer from Celeb-DF). This is the deployment-level manifestation of the artefact-learning problem: high in-distribution AUC coexisting with near-chance out-of-distribution fake recall.

Implications for v2: These causes suggest concrete, mostly training-free fixes, in order of expected impact: (i) replace the fixed $\tau = 0.5$ with a prevalence- or validation-calibrated threshold; (ii) widen or remove the VLM shift-clamp and route on joint low-confidence (not only disagreement) so out-of-distribution agreements are also corrected; (iii) retrain the ensemble on a generator-diverse corpus that

includes diffusion-based forgeries to close the train/test family gap; and (iv) add manipulation-sensitive supervision (frequency-domain or blending-boundary cues) so the representation encodes artefacts rather than authenticity alone.

7. Conclusion

This paper presented a unified framework for the ImageCLEF 2026 Deepfake Detection and Generation Task across both image and audio modalities. The experimental results demonstrated strong performance in authentic-content recognition, speech intelligibility, and preservation of speaker similarity. However, the framework showed reduced robustness against highly realistic and previously unseen synthetic manipulations, particularly in image deepfake detection and spoofing-evasion scenarios. Overall, the findings highlight the challenges of generalizable deepfake analysis and suggest that future improvements should focus on out-of-distribution robustness, stronger identity preservation, and enhanced adversarially-aware training strategies.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools solely for grammar correction, spell-checking, and minor language polishing in order to improve the readability and clarity of the manuscript. After using these tools, the authors carefully reviewed and edited the resulting text and take full responsibility for the content of this publication.

Acknowledgments

This work used computational resources provided by the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program under Allocation Request **CIS251242**, including **53.0 GPU hours on the NCSA Delta GPU system**. The author acknowledges the National Science Foundation (NSF), ACCESS, and NCSA Delta for supporting the computational resources used in this research.

8. Online Resources

GitHub: <https://github.com/Ejigsonpeter/Constraint-Guided-Generation-for-Audio-Visual-Deepfake-Evaluation>

References

- [1] I. Amerini, M. Barni, P. Bestagini, et al., Deepfake media forensics: Status and future challenges, *Journal of Imaging* 11 (2025) 73.
- [2] G. Pei, et al., Deepfake generation and detection: A benchmark and survey, *ACM Computing Surveys* 58 (2026). Early access / in press.
- [3] M. Hoque, F. Kim, Generative modeling of additive manufacturing lack-of-fusion defects using three-dimensional adversarial network (2026).
- [4] L. Meng, et al., Audio deepfake detection: What has been achieved and what remains to be explored, *Sensors* 25 (2025) 1812.
- [5] 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.
- [6] 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.
- [7] A. Raza, A. Basit, A. Amin, Z. A. Arfeen, M. I. Masud, U. Fayyaz, T. A. Jumani, A comprehensive review of deepfake detection techniques: From traditional machine learning to advanced deep learning architectures, *AI* 7 (2026) 68. doi:10.3390/ai7020068.
- [8] R. Kundu, H. Zhou, A. K. Roy-Chowdhury, et al., Towards a universal synthetic video detector: From face or background, in: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 1–10.
- [9] R. N. Chowdhury, M. Hoque, M. R. Hasan, E. P. O. Oluwafemi, M. M. Rahman, Modality-guided radiology caption prediction with small vision-language models and image classifier, in: *CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org*, Madrid, Spain, 2025.
- [10] M. Hoque, M. Hasan, M. Emon, F. Khalifa, M. Rahman, Medical image interpretation with large multimodal models, in: *CEUR workshop proceedings, CEUR Workshop Proceedings 3740, CEUR-WS. org* 2024, 2024.
- [11] Q. Zhang, S. Wen, T. Hu, et al., Audio deepfake detection with self-supervised XLS-R and SLS classifier, in: *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2024, pp. 1–9.
- [12] N. Foteinopoulou, et al., A hitchhiker’s guide to fine-grained face forgery detection using common sense reasoning, in: *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024, pp. 1–14.
- [13] M. Hoque, R. N. Chowdhury, M. R. Hasan, O. O. E. Peter, F. Khalifa, M. M. Rahman, An empirical evaluation of low-rank adapted vision–language models for radiology image captioning, *Bioengineering* 12 (2025) 1330.
- [14] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, et al., XTTS: A massively multilingual zero-shot text-to-speech model, in: *Proc. Interspeech*, 2024, pp. 4978–4982.
- [15] O. O. Ejiga Peter, M. M. Rahman, F. Khalifa, Advancing AI-powered medical image synthesis: Insights from MedVQA-GI challenge using CLIP, fine-tuned stable diffusion, and DreamBooth + LoRA, *arXiv preprint arXiv:2502.20667*, 2025. <https://arxiv.org/abs/2502.20667>.
- [16] O. O. Ejiga Peter, O. G. Akingbola, C. R. Amalahu, O. Adeniran, F. Khalifa, M. M. Rahman, Synthetic data-driven multi-architecture framework for automated polyp segmentation through integrated detection and mask generation, *Medical Imaging 2025: Clinical and Biomedical Imaging, SPIE* (2025) 78. doi:10.1117/12.3049369.
- [17] 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: *Proc. Interspeech*, 2024, pp. 4978–4982.
- [18] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, InstantID: Zero-shot identity-preserving generation in seconds, *arXiv preprint arXiv:2401.07519*, 2024. <https://arxiv.org/abs/2401.07519>.
- [19] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A large-scale challenging dataset for DeepFake forensics, in: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3207–3216.
- [20] Qwen Team, Qwen2.5-vl technical report, *arXiv preprint arXiv:2502.13923*, 2025. <https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct>.
- [21] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baeovski, A. Conneau, M. Auli, XLS-R: Self-supervised cross-lingual speech representation learning at scale, in: *Proc. Interspeech*, 2022, pp. 2278–2282.