

Overview of ImageCLEF 2026 Deepfake Task: Multimodal Detection and Generation of Deepfakes

Dan Cristian Stanciu^{1*}, Andrei Radu¹, Radu George Bolborici¹, Marian Negru¹, Alexandru Florin Ene¹, Vlad Mihai Vasilescu¹, Ana Antonia Nicolae¹, Liviu Daniel Ștefan¹, Mihai Gabriel Constantin¹, Mihai Dogariu¹, Alexandra Georgiana Andrei¹ and Bogdan Ionescu¹

¹National University of Science and Technology Politehnica Bucharest, Romania

Abstract

The first edition of the ImageCLEF Deepfake Task aims to advance research in deepfake generation and detection by evaluating methods under realistic and challenging conditions. The task was organized across two modalities, Images and Audio, and included two interconnected subtasks: Deepfake Generation and Deepfake Detection. In the Generation subtask, participants created synthetic media from real data while satisfying identity and content constraints, with generated samples subsequently evaluated for perceptual quality and their ability to evade detection systems. In the Detection subtask, participants developed binary classification models to distinguish between authentic and manipulated content using datasets containing both organizer-generated and participant-generated deepfakes. A total of 67 groups registered for the Generation task and 90 for the Detection task, resulting in 118 submitted runs for Generation and 87 for Detection. 22 teams had valid submissions across the subtasks, resulting in 14 received working notes papers. Results showed that high-quality deepfakes were often more successful at evading detectors, highlighting the growing realism of modern generative models. In audio detection, several approaches demonstrated strong generalization and achieved high detection accuracy. In contrast, image detection remained substantially more difficult, with many systems achieving only modest improvements over chance-level performance. Overall, the task provided valuable insights into the challenges of building robust deepfake detection systems and emphasized the need for methods capable of generalizing across diverse generation techniques and unseen manipulations.

Keywords

Deepfake, Multimodal, Audio, Images, Generalization, Deepfake Detection, Deepfake Generation

1. Introduction

Deepfakes are defined as synthetic media generated using artificial intelligence to produce realistic representations of real individuals in images, audio, or video. As generative models have advanced in recent years, the visual and perceptual quality of deepfakes has improved substantially, making them increasingly difficult for humans and automated systems to distinguish from authentic content. While these technologies have enabled a wide range of beneficial applications, including film production, virtual avatars, accessibility tools, education, and privacy-preserving content generation, they have also introduced significant societal challenges.

The widespread availability of deepfake generation tools has facilitated the creation of convincing manipulated media that can be used for misinformation, identity fraud, impersonation, financial scams, and non-consensual content. Such misuse poses risks to public trust, digital security, and the integrity of online information ecosystems. Consequently, the development of reliable deepfake detection and localization methods has become an important research area, aiming to mitigate these risks while supporting the responsible use of generative artificial intelligence.

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

*Corresponding author.

✉ dan.stanciu1203@upb.ro (D. C. Stanciu)

ORCID 0009-0008-4561-4328 (D. C. Stanciu); 0009-0008-4774-373X (A. Radu); 0009-0009-3646-6519 (R. G. Bolborici); 0009-0005-7665-139X (M. Negru); 0009-0002-7129-3018 (A. F. Ene); 0000-0003-3743-4235 (V. M. Vasilescu); 0009-0003-4624-9442 (A. A. Nicolae); 0000-0001-9174-3923 (L. D. Ștefan); 0000-0002-2312-6672 (M. G. Constantin); 0000-0002-8189-8566 (M. Dogariu); 0009-0003-5649-3078 (A. G. Andrei); 0000-0003-4112-5769 (B. Ionescu)



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

Nowadays, the biggest problem surrounding deepfakes is the lack of generalization in detector models. While deepfake detectors can be trained on a multitude of datasets, the final model may learn only features from those specific datasets, thus failing in real-life situations. Therefore, it is essential to continue research activity on ways to increase generalization and protect the public from the threat of deepfakes.

The ImageCLEF 2026 [1] Deepfake task aims to bridge the gap between implemented models and real-life settings in deepfakes by introducing two complementary subtasks, aimed at finding ways to make deepfake detectors generalize and studying what makes a deepfake hard to detect. The two tasks are: Deepfake Generation and Deepfake Detection.

The Generation task challenges participants to create deepfakes in two modalities: Images and Audio. The generation process is subject to a set of constraints designed to ensure that the synthesized content is both realistic and representative of real-world manipulations, while simultaneously maximizing its ability to evade detection. Evaluating the robustness of deepfake detectors against such challenging examples is essential for understanding the strengths and limitations of current detection methods. Consequently, the generated deepfakes are assessed not only for their perceptual realism but also for their effectiveness in deceiving detection models, as they are incorporated into the evaluation dataset for the Detection task.

The Detection task is closely interconnected with the previously described Generation task, as the media generated by participating systems, together with additional authentic and manipulated samples, serves as input to the detection models. Since the deepfakes are produced using a diverse range of generation techniques, many of which differ from those represented in existing benchmark datasets, conventional detection approaches may struggle to generalize across all manipulation methods. Consequently, participants are encouraged to develop novel and robust detection strategies that improve generalization to previously unseen deepfake generation techniques, enabling reliable performance under realistic conditions.

The main contributions of this paper are as follows: (1) we introduce the deepfake generalization problem and define two complementary tasks designed to investigate the factors that make deepfakes difficult to detect and enable detectors to generalize to unseen manipulations; (2) we present the datasets and evaluation protocols developed for two modalities, Images and Audio; (3) we analyze and discuss the results obtained by the 22 participating teams; (4) we summarize key insights on detector generalization derived from the participants' technical reports and the experimental results; and (5) we discuss the main conclusions of the challenge, its current limitations, and potential directions for future editions.

2. Task and Participation

2.1. Task Description

The Deepfake task is split into two sub-tasks, which are strongly interconnected, with the results from one subtask being used in the evaluation of the other and vice versa. Each sub-task with two independent modalities: **audio** and **images**. The sub-tasks are:

- **Deepfake Generation:** The generation task asked participants to generate synthetic data under some specific constraints. The generated data should be as realistic as possible and also evade deepfake detection at the same time. This is a few-shot problem, with both modalities providing a limited amount of data to guide generators. For **Images**, the organizers provided a dataset of videos of short length (15-45 seconds), one per identity. The participants were tasked with generating frames containing the same identity, with real frames being used as a target. To ensure a correct generation of the ground truth (GT) frames, the participants were provided with a set of facial landmarks extracted from the GT frames to be used as a target. Another restriction was the preservation of identity: participants were tasked to provide the required identity, and embeddings extracted from their output frames were compared to an averaged embedding of the person's identity. For **Audio**, the organizers provided a set of 15 audio-text pairs for each out

of 16 different identities. The participants were asked to generate new audio files of a person saying specific phrases, while at the same time preserving audio quality and the identity. Mode, for both modalities, the generated data was given as input in the Detection task. The average of the accuracy obtained in the Detection task on each submission was calculated, and the Evasion score was computed as $1 - \text{Detection_Accuracy}$. A final Overall score was calculated using the Quality score and the Evasion score.

- **Deepfake Detection:** The Detection task is defined as a binary classification problem across both modalities, Images and Audio. Participants were required to classify each image or audio file as either Real (0) or Deepfake (1). For both modalities, the evaluation data comprised Real data (the real data from the Generation task together with a secret set of videos featuring the same identities) and Deepfake data (deepfakes produced by participants in the Generation task and deepfakes generated by the organizers). The final score was computed as a weighted accuracy over these four data types. A notable aspect of the Detection task is that no training data was provided. This decision was motivated by two considerations. First, a substantial amount of deepfake and generated image data is already available in the literature. Second, and more importantly, providing training data would undermine the purpose of the experiments: it is well established that most detection models fail to generalize, instead learning dataset-specific artifacts rather than features indicative of manipulation.

Figure 1 presents the ImageCLEF 2026 tasks, computed metrics and how they interact with each other.

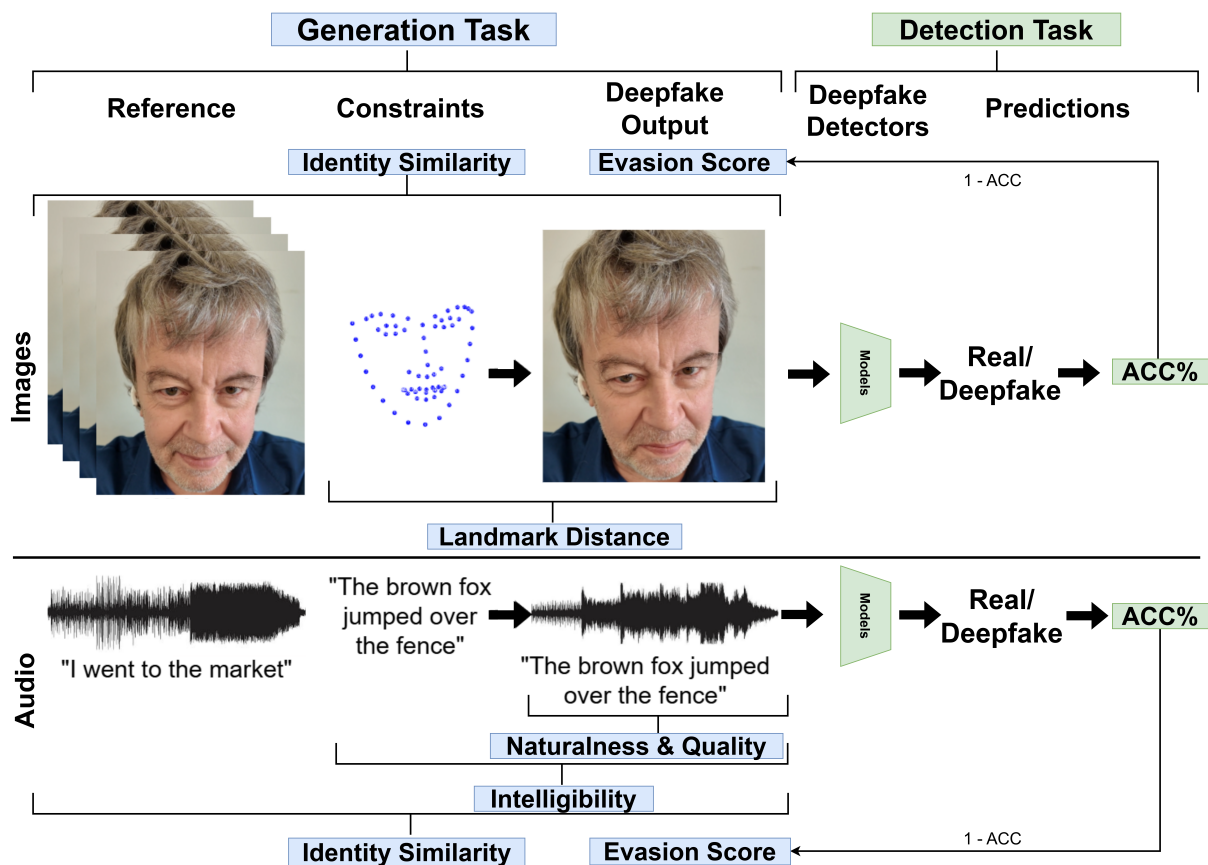


Figure 1: Diagram presenting the ImageCLEF 2026 Deepfake Tasks, the metrics computed for each modality and how they interact with each other.

2.2. Participation

The tasks were hosted on the Ai4Media benchmarking platform ^{1,2} for automatic Evaluation, Data hosting and participant management. The platform allowed participants easy access to the datasets and performed automatic results assessment and scoring.

Due to the nature of the evaluation, the results from the Detection task were required to evaluate the Generation task, and the data from the Generation task was used as input in the Detection task. Therefore, the tasks were executed sequentially, with the Generation task running for a month and the Detection task running for the next one. Some metrics were calculated retroactively for both tasks. Further details regarding the evaluation will be presented in Chapter 4.

Submissions were allowed on a per-user basis, with the results being aggregated after the competition. Each user had a maximum of 10 possible submissions. Similar or identical submissions were aggregated under the same team. For the Generation task, we selected the best submission of the last 3 sent by the team. For the Detection task, each team was instructed that only the last submission would be considered for the final results.

For the **Generation** Task, 10 teams submitted for the Audio modality and 11 teams submitted for the Image modality. Of these, 10 teams participated in both modalities, showing great synergy and interest in all modalities for generation.

For the **Detection** task, 15 teams participated in the Audio modality and 22 teams participated for Images. Again, the synergy between the tasks was strong: all the teams that participated in the Audio modality also participated in Images. Furthermore, all the teams that participated in the Generation task also participated in the Detection task.

A total of 67 users registered for the Generation task and 90 for the Detection task, resulting in 118 submitted runs for Generation and 87 for Detection. A total of 22 different teams (26 users) submitted results for both tasks, with all the teams in the Generation task also participating in the Detection task. A total of 205 runs were submitted for both tasks and 15 working notes papers were submitted, with 13 of them being accepted for presentation at CLEF.

3. Data

For the Generation task, we provided two datasets for the Images and Audio modality, containing real data, to be used to generate deepfakes. For Audio, 15 audio files of different lengths were provided, for a total of 16 speakers. For Images, we provided 335 videos of short length displaying a person talking. Both tasks are designed as few-shot problems, only having a limited number of data points for each identity.

For the Detection task, for both modalities, the participants received a set aimed at binary classification of deepfakes, comprised of (i) the real data provided in the Generation Task, (ii) a set of real data that was kept out of the generation task, displaying the same identities, (iii) organizer-generated deepfakes, using multiple different methods, and (iv) the deepfakes generated by participants in the Generation task. For the generation task, **no training data was provided**, as this defeats the purpose of studying generalization in deepfake detection - models usually learn dataset-specific features, not generalizable features.

The data for both subtasks was made available on the Ai4Media platform to all accepted participants in the competition.

3.1. Audio

The audio data component of the Deepfake task consisted of two related collections: (i) a generation dataset for speaker-conditioned audio synthesis and (ii) a detection dataset for real-versus-fake audio classification. In the generation dataset, participants received real reference recordings alongside the

¹<https://ai4media-bench.aimultimedialab.ro/competitions/14/>

²<https://ai4media-bench.aimultimedialab.ro/competitions/13/>

associated transcripts for each speaker, together with target transcripts for which new audio had to be synthesized. The corresponding real recordings for the target transcripts were withheld from participants and used only as hidden ground truth during the official evaluation.

The public audio generation dataset comprised of 16 speaker identities recorded by the organizers, including 11 male and 5 female speakers aged between 22 and 32. All utterances were spoken in English, while the speaker set included variation in accent, speaking style, and recording conditions. The released reference material consisted of speech clips, between 2 and 10 seconds long, distributed as 16-bit monophonic WAV files. The utterances were divided into three transcript-length categories, denoted as *Short*, *Medium*, and *Long*, according to the number of characters in the transcript. Fig. 2 provides representative transcript examples for each of the three utterance-length categories.

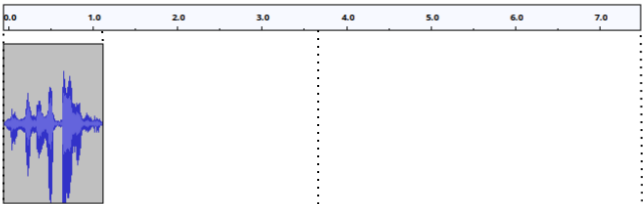
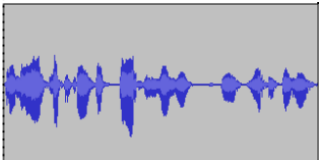
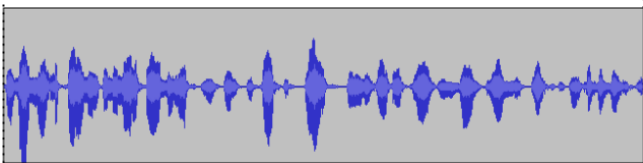
Category	Representative transcript	Audio
Short	"But this is a fallacy."	
Medium	"So they went up across the square and wondered still more as they went."	
Long	"He had no occasion to delay for at the next instant a burst of cries filled the outer air and ran along the whole extent of the village."	

Figure 2: Representative transcript examples for the three utterance-length categories used in the audio generation data.

For each speaker, the reference set contained 30 real audio-text pairs: 15 *Short* utterances, 10 *Medium* utterances, and 5 *Long* utterances. Across all 16 speakers, this resulted in 480 released reference recordings. The generation target set followed the same speaker identities and transcript-length categories, but contained only text prompts for which participants had to submit synthesized speech of. Each speaker was associated with 30 target utterances: 5 *Short*, 10 *Medium*, and 15 *Long*. Consequently, the audio generation subtask required 480 synthesized files in total. The metadata associated with the public data linked each item to its speaker identity, transcript-length category, and textual content, allowing generated samples to be evaluated against the intended speaker and phrase. Submitted audio was required to be provided as 16-bit, mono WAV files, sampled at 16 kHz, ensuring a uniform audio format for evaluation.

The audio detection dataset was constructed as a binary authenticity classification collection, with samples labeled as either real or deepfake speech. The real class included authentic recordings from the public audio reference set, together with additional authentic recordings from a private set that was not released to the participants. The deepfake class combined organizer-generated synthetic speech with participant-generated submissions from the audio generation subtask. As a result, the fake class was not tied to a single synthesis pipeline, but instead included outputs produced under different generation conditions and by different systems. This composition was intended to support the evaluation of more general audio deepfake detectors, rather than models specialized for artifacts of a single generator.

For obtaining the organizer-generated synthetic speech samples, three zero-shot TTS models were

used: XTTSv2 [2], CosyVoice3 [3] and IndexTTS2 [4], using a single randomly selected reference utterance from the target speaker. Additionally, a fine-tuning-based approach, namely UnitSpeech [5], was used. Consequently, one set of weights was trained for each speaker using all 30 available references. Inference was applied with all models following the audio deepfake generation submission format.

In total, the audio detection dataset contained 11,520 files, comprising 960 real recordings and 10,560 deepfake samples. The integration of participant-generated samples directly linked the generation and detection subtasks: submitted voice-cloning outputs were evaluated not only as synthesized utterances, but also as part of the detection benchmark. This design produced a heterogeneous detection dataset, reflecting the practical requirement that audio deepfake detectors should operate across diverse and potentially unseen synthesis methods.

3.2. Images

For the **Generation** task, participants received 335 development videos of length between 15 and 45 seconds, each corresponding to a different identity. The videos were extracted from Youtube clips and had CC-BY license. The data was selected from longer videos containing a single person speaking, with their face visible and close to the camera. The videos depict people facing the camera, in a variety of angles, lighting conditions, qualities. Moreover, the dataset is extremely diverse, capturing individuals from all races, cultures, with a balance in age and gender. Different parts of the same videos were held out as a secret dataset, used for evaluation. A document containing video sources was provided along with the videos.

The participants also received a set of 1-3 sets of 478 facial landmarks extracted using the mediapipe python library, along with a few optional pose landmarks extracted with the same library. The landmarks were extracted from frames contained in the secret dataset. In total, 1002 identity-landmark set pairs were generated, to be used as targets for generation. Participants were tasked with generating an image of size 256x256, while preserving the identity and the landmark positions.

For the **Detection** task, a total of 24,404 RGB images of size 256x256 in .png format were provided. The dataset was comprised of (i) 1,002 real frames from the development set, (ii) 3,400 real frames extracted from the secret set, including the 1,002 frames that were used as source for generating the landmarks constraints, (iii) 3,970 deepfakes generated by the organizers using 4 different generators, (iv) 16,032 deepfakes generated by the other participants in the Generation task.

The organizer-generated deepfakes were generated to be used as a baseline for the performance of the models submitted by participants. The dataset was designed to be simple, straight forward, open-source methods that could be used by participants in this task. Each method was used to generate slightly different images, but also used different random frames as an input. This is used to test whether the detectors would overfit and learn the facial expressions and body positions instead of useful features. The methods used are:

- The first method used was original Stable Diffusion v1.5 [6] latent diffusion model, applied via the image-to-image pipeline. Real frames are transformed under a low denoising strength of 0.35, guidance scale 7.5, and 40 inference steps. The low strength preserves the original identity and composition while introducing subtle diffusion-model artifacts, producing conservative forgeries close to the real distribution. As the standard, widely benchmarked model, it serves as the reference point against which the more specialized generators are compared. The model used a positive/negative prompt of "photo of the same person speaking, realistic portrait, natural lighting, slight mouth open, moving hands" and "cartoon, painting, blurry, distorted face, extra eyes, bad anatomy".
- The second generator replaces the base checkpoint with Photon v1 [7], a community fine-tune of Stable Diffusion v1.5 optimized for photorealistic portraits. The pipeline configuration is otherwise identical to Model 1 (strength 0.35, guidance 7.5, 40 steps), isolating the effect of the fine-tuned weights. It was included to test whether a photorealism-tuned model within the same

architecture yields more convincing skin texture and lighting, adding intra-architecture diversity without changing generation dynamics. This generator used the same prompt as the first one.

- The third generator moves to the larger SDXL [8] architecture, using the RealVisXL V4.0 [9] checkpoint at 1024×1024 native resolution. Relative to the SD 1.5 models it uses a higher denoising strength (0.55) and a per-image random seed, granting greater pose variation and higher-fidelity output. It was chosen to introduce a distinct, higher-capacity architecture whose forgeries differ more substantially from the source frames, stress-testing detectors against modern high-resolution generators. The images were reduced in size, to 256×256, to fit the required size. This model drops the explicit mouth and hand cues in favor of a "different head position" cue, consistent with its higher denoising strength and goal of greater pose variation
- The last model was an implemented Landmark-Conditioned Autoencoder. The convolutional autoencoder is trained from scratch to synthesize faces conditioned jointly on identity and facial geometry. It reconstructs a target face from three signals: a reference image of the identity, a rendered landmark-point image, and the raw MediaPipe FaceMesh coordinates. This model was used to establish a different baseline and due to the fact that autoencoders might produce different fingerprints, as opposed to Stable Diffusion. This model also used images from the secret test in training.

Figure 3 presents a comparison of frames from four identities included in the Detection dataset, with the labels on the left indicating the origin of each image. A key aspect of the dataset is the separation of authentic data into two subsets: the development set, which is available to participants during model development, and the secret set, which is reserved exclusively for evaluation. The secret samples are typically extracted from different segments of the same source videos, with the goal of introducing variations in pose, camera distance, lighting, or facial expression. For example, the first identity exhibits a noticeable change in viewpoint and illumination, with the subject appearing closer to the camera in the secret sample. In contrast, the middle examples illustrate that, in some cases, the visual differences between the development and secret sets are relatively minor.

The generated samples further highlight the diversity of manipulation methods. The most apparent difference is the variation in identity preservation across generation techniques. The Stable Diffusion-based models intentionally introduce semantic modifications, resulting in synthesized faces that do not always faithfully preserve the original identity. Among these methods, the baseline Stable Diffusion 1.5 model exhibits visible visual artifacts, whereas the more recent diffusion models produce images with substantially higher perceptual realism. In contrast, the autoencoder-based approach aims to preserve the original frame as closely as possible, although it may introduce characteristic convolutional artifacts and a slight loss of sharpness.

An important aspect of the participant-generated images is that they were compared for similarity to ensure that different participants did not submit identical or near-identical samples. This verification step helped maintain diversity among the submitted deepfakes and prevented the evaluation from being influenced by highly similar or duplicated content. Figure 4 presents examples of frames generated by some of the the participating teams, along with the Target image from the secret set. The participant-generated images have a quality score attributed to them, which will be explained in the next chapter. The difference in perceived quality can be easily observed in the comparison. The final dataset was composed of data of different quality levels, challenging the participating teams to adapt their models.

4. Evaluation Methodology

4.1. Audio

The audio evaluation methodology comprised two components, corresponding to the generation and detection subtasks. In the generation subtask, submitted waveforms were evaluated according to speech quality and detection evasion. In the detection subtask, submitted systems were evaluated as binary classifiers distinguishing authentic speech from synthetic one across multiple audio sources.

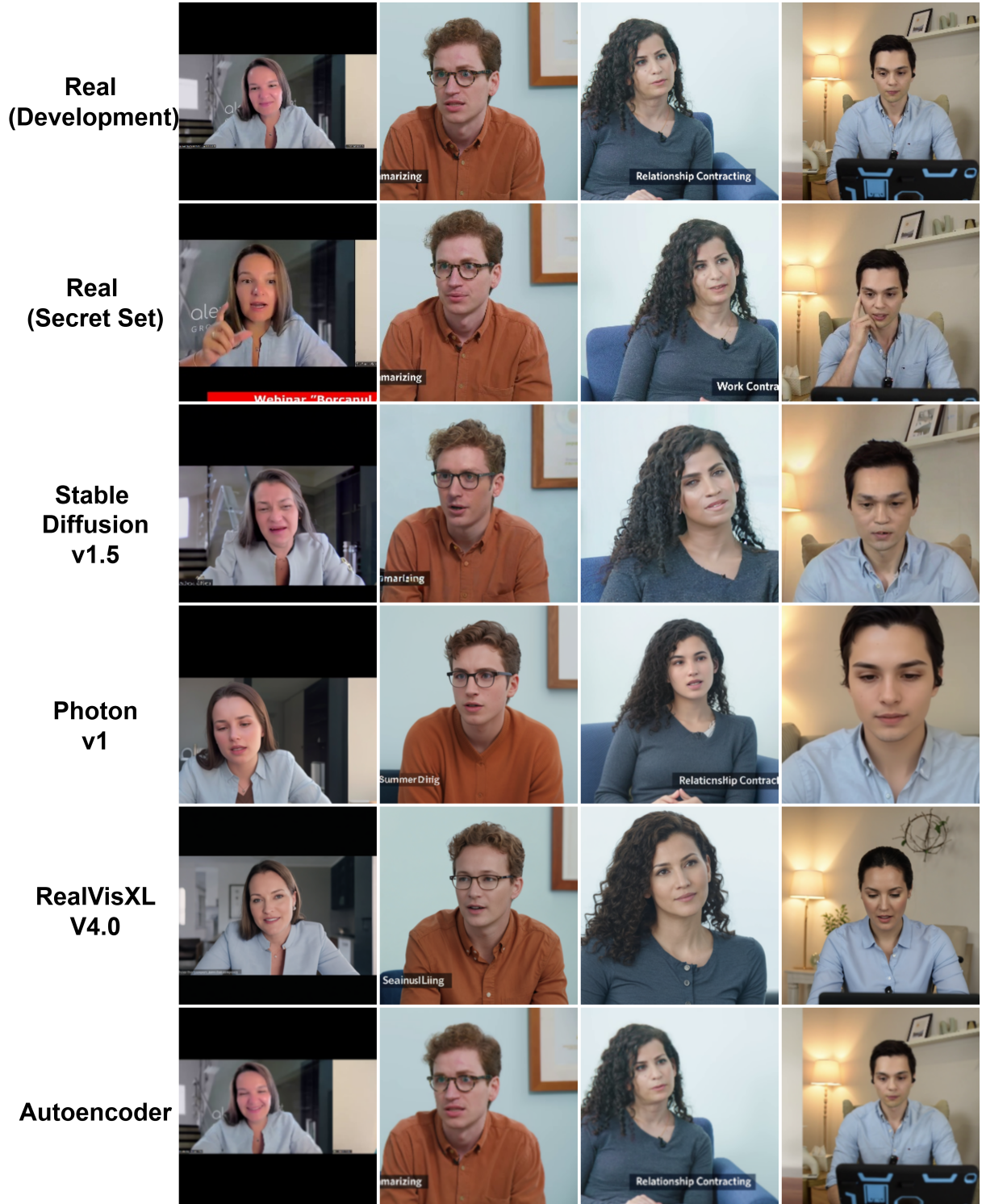


Figure 3: Comparison of organizer-provided frames in the Image Detection Subtask. The organizers provided real frames from the development videos, real frames from the secret set and data generated using 4 different models, based on development data. The text to the left specifies the images provenience.

For audio generation, each submitted utterance was evaluated against the hidden ground-truth recording of the same speaker reading the same target transcript. The quality component combined speaker similarity, intelligibility, and naturalness. Speaker similarity was measured by computing the speaker embedding cosine similarity (SECS) between the generated and ground-truth recordings, using both ECAPA-TDNN [10] and WavLM-based [11] speaker encoders. Intelligibility was measured by

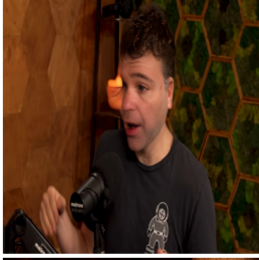
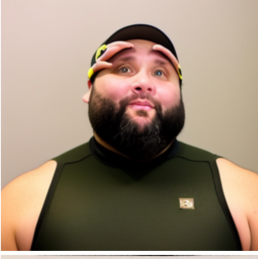
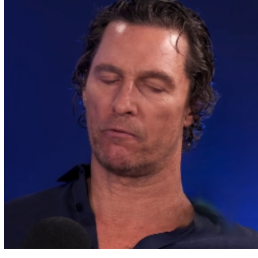
Quality Score \ ID	0	1	2	3
Target				
0.7				
0.61				
0.46				
0.37				
0.25				

Figure 4: Comparison of frames generated by participants in the Image Detection Subtask, based on Image quality score, for the first four identities.

Table 1

Normalization of the audio generation quality metrics.

Metric	Effective interval	Score used
NISQA MOS	[1, 5]	$S_{\text{NISQA}} = \frac{\text{MOS}-1}{4}$
ECAPA SECS	[0, 1]	$S_{\text{ECAPA}} = \text{SECS}_{\text{ECAPA}}$
WavLM SECS	[0, 1]	$S_{\text{WavLM}} = \text{SECS}_{\text{WavLM}}$
WER	[0, ∞)	$S_{\text{WER}} = 1 - \min(\text{WER}, 1)$
CER	[0, ∞)	$S_{\text{CER}} = 1 - \min(\text{CER}, 1)$

transcribing the generated audio with Whisper large-v3 [12] and comparing the resulting transcription with the target transcript using Word Error Rate (WER) and Character Error Rate (CER). Naturalness and perceptual quality were measured using NISQA Mean Opinion Score (MOS) [13], a non-intrusive estimator of perceived speech quality. Before aggregation, all metrics were converted so that higher values indicate better performance and all quality components lie in the $[0, 1]$ interval. The NISQA MOS score is originally defined on a $[1, 5]$ scale and was normalized as $S_{\text{NISQA}} = (\text{MOS} - 1)/4$. The ECAPA-TDNN and WavLM speaker-similarity scores were cosine-similarity-based SECS values used as speaker-identity preservation scores, with higher values indicating greater similarity. For the speaker-similarity metrics, the SECS values were used directly. Although cosine similarity is theoretically defined over $[-1, 1]$, the evaluated same-speaker speaker-embedding comparisons produced non-negative values in practice; therefore, ECAPA-TDNN and WavLM SECS were treated as scores in the effective $[0, 1]$ interval and were not further normalized. WER and CER are error rates, where lower values indicate better transcript preservation; therefore, they were converted into accuracy-like scores as $S_{\text{WER}} = 1 - \min(\text{WER}, 1)$ and $S_{\text{CER}} = 1 - \min(\text{CER}, 1)$. Table 1 summarizes the effective metric intervals and the transformations used before computing the audio quality score. After this step, all components are expressed so that larger values indicate better generation quality.

The audio quality score was computed as a weighted aggregation of normalized metric values:

$$Q_{\text{audio}} = 0.40S_{\text{NISQA}} + 0.15S_{\text{ECAPA}} + 0.15S_{\text{WavLM}} + 0.15S_{\text{WER}} + 0.15S_{\text{CER}}, \quad (1)$$

where S_{NISQA} denotes the normalized NISQA MOS score, S_{ECAPA} and S_{WavLM} denote the speaker-similarity scores obtained with the two speaker encoders, and S_{WER} and S_{CER} denote the transcript-preservation scores derived from WER and CER, respectively. The weights were chosen to reflect the intended priorities of the audio generation evaluation. NISQA receives the largest weight, 0.40, because perceptual naturalness and speech quality are central for the usefulness of a generated utterance. Speaker identity preservation is measured with two complementary encoders, ECAPA-TDNN and WavLM, each weighted with 0.15 in order to reduce dependence on a single speaker-representation model. Transcript preservation is represented by both WER and CER, also weighted with 0.15 each, balancing word-level and character-level intelligibility.

Generation submissions were also evaluated by their ability to evade deepfake detectors. The evasion score, E_{audio} , combined the average evasion against participant-submitted detection systems and organizer-side detectors:

$$E_{\text{audio}} = 0.70 E_{\text{participants}} + 0.30 E_{\text{organizers}}, \quad (2)$$

where $E_{\text{participants}}$ denotes the evasion score computed using participant detection systems, and $E_{\text{organizers}}$ denotes the evasion score computed using SSL-AASIST [14], the organizer-side detector. For the evasion component, a larger weight is assigned to participant-submitted detectors because they represent the main competitive detection setting and include a broader range of detection strategies. The organizer-side detector is weighted lower but kept in the score as a fixed reference system, ensuring that generation submissions are also evaluated against a stable baseline detector.

The final audio generation score was computed as:

$$S_{\text{gen,audio}} = Q_{\text{audio}} \cdot E_{\text{audio}}. \quad (3)$$

This formulation ranked systems according to both the quality of the synthesized speech and its resistance to detection.

For the audio detection subtask, participants submitted binary authenticity predictions for audio samples labeled as real or deepfake. The evaluation considered four categories: organizer-generated deepfakes, real samples from the public reference set, real samples from the hidden ground-truth set, and participant-generated deepfakes. The final audio detection score was computed as a weighted average of category-level accuracies:

$$S_{\text{det,audio}} = 0.10 A_{\text{orgFake}} + 0.10 A_{\text{refReal}} + 0.40 A_{\text{gtReal}} + 0.40 A_{\text{partFake}}, \quad (4)$$

where A_{orgFake} denotes accuracy on organizer-generated deepfakes, A_{refReal} denotes accuracy on real reference recordings, A_{gtReal} denotes accuracy on hidden real recordings, and A_{partFake} denotes the weighted accuracy on participant-generated deepfakes. This category-level scoring reduced the effect of class imbalance and emphasized generalization to diverse synthetic speech sources rather than performance on a single generation method. The detection weights emphasize the subsets most relevant to generalization. Hidden real recordings and participant-generated deepfakes receive the largest weights, 0.40 each, because they test performance on unseen authentic data and on diverse synthetic samples produced by different submitted generation systems. Organizer-generated deepfakes and public reference real recordings receive lower weights, 0.10 each, because they serve primarily as anchor subsets for controlled evaluation while being less representative of the main hidden generalization challenge.

4.2. Images

4.2.1. Image Generation Metrics

The image generation sub-task is evaluated along two complementary axes: structural fidelity of the face and preservation of identity. Two raw metrics are computed for each submission by comparing the generated image against a reference drawn from the corresponding real video.

The first metric, *Landmark Distance*, is the mean Euclidean distance between the facial landmarks of the generated image and those of the reference, and captures how well the structural geometry of the face is preserved; lower values are better. The second metric, *Identity Similarity*, is the cosine distance between face-identity embeddings of the generated image and a reference obtained by averaging over all frames of the corresponding real video.

Because the two raw metrics are on different and unbounded scales, each is normalized to the unit interval $[0, 1]$, with higher always meaning better, before being combined. The Landmark Distance is mapped through a concave, non-linear transform anchored to a baseline of $b = 0.137089$, the mean landmark distance between randomly matched faces, which represents the worst-case behavior of a system producing unrelated faces, calculated by the organizers. The score is defined as $\max(0, 1 - (d/b)^{0.5})$, where the square-root exponent $\alpha = 0.5$ gives the curve its concave shape: improvements among already-strong submissions (small distances) are rewarded more generously than equivalent gains at the high-distance end. A perfect submission with zero landmark distance receives a score of 1.0, while any submission at or beyond the baseline is clipped to 0.0.

For the Identity Similarity, the ground-truth frames themselves exhibit a mean self-distance of approximately 0.7, which we adopt as the upper anchor of the “good” range, while 0.4 marks the lower bound below which a generation is treated as failing to preserve identity. The Identity Similarity is therefore linearly rescaled between these two bounds, $\frac{d_{\text{ID}} - 0.4}{0.3}$, with values of 0.7 or above mapped to 1.0 and values of 0.4 or below mapped to 0.0; higher scores indicate identity better aligned with the ground-truth reference. The overall **Image Quality Score** is the unweighted mean of the two normalized scores.

The Image Quality Score alone rewards visually faithful generation, but the objective of the deepfake generation task is to produce images that are both realistic *and* difficult to detect. We therefore combined quality with two evasion measures. The *Deepfake Evasion Score (Participants)* quantifies how well a

Table 2

Normalization of the image generation metrics and composition of the final score.

Metric	Effective interval	Score used
<i>Quality metrics</i>		
Landmark Distance	$[0, b], b = 0.137089$	$S_{\text{LM}} = \max(0, 1 - (d_{\text{LM}}/b)^{0.5})$
Identity Similarity	$[0.4, 0.7]$	$S_{\text{ID}} = \min(1, \max(0, \frac{d_{\text{ID}} - 0.4}{0.3}))$
Image Quality	$[0, 1]$	$S_{\text{IQ}} = \frac{1}{2} S_{\text{LM}} + \frac{1}{2} S_{\text{ID}}$
<i>Final score</i>		
Evasion (Participants)	$[0, 1]$	$E_{\text{part}} = 1 - \overline{\text{acc}}_{\text{part}}$
Evasion (Organizers)	$[0, 1]$	$E_{\text{org}} = 1 - \overline{\text{acc}}_{\text{org}}$
Final Score	$[0, 1]$	$S_{\text{final}} = S_{\text{IQ}} (0.7 E_{\text{part}} + 0.3 E_{\text{org}})$

submission escapes the detection systems contributed by other participants, computed as one minus the mean detection accuracy across those systems; the *Deepfake Evasion Score (Organizers)* is defined identically but against the organizers’ own detectors, providing a stable and consistent reference point across all submissions. In both cases, higher values indicate harder-to-detect generations.

The organizer-implemented evasion score is computed against a fixed ensemble of open-source deepfake detectors, namely from the DeepfakeBench [15] project, which provides standardized implementations and pre-trained weights for a broad range of published detection methods. All the detectors were trained on FaceForensics++ [16], which provides simple baseline models trained on a state-of-the-art dataset. We employ the twelve RGB detectors that require no auxiliary inputs such as facial landmarks or segmentation masks, namely Xception [17], EfficientNet-B4 [18], Meso4 [19], Meso4-Inception [19], F3Net [20], SPSL [21], SRM [22], UCF [23], CORE [24], and RECCE [25]. We also used two organizer implemented models using an Xception [17] architecture, trained recursively using adversarial attacks [26] and trained with autoencoder-based augmentations [27]. The aforementioned models were trained on FaceForensics++ and CelebDF [28]. Each submitted image is resized to 256×256 and normalized before being passed through every detector, which returns a probability that the image is fake. A detector flags an image as fake when this probability is at least 0.5; per detector we record both the continuous probability and the binary decision. For each detector we compute its accuracy as the fraction of a submission’s generated images correctly identified as fake, and the organizer evasion score E_{org} is then defined as one minus the mean accuracy across the ten detectors, so that a submission whose images consistently fool the ensemble receives a high evasion score. Because these detectors are fixed, publicly available, and applied identically to every submission, they provide a controlled and reproducible reference point that complements the participant-contributed detectors, whose composition varies with the field of submissions.

The **Final Score**, which serves as the primary ranking metric, is the product of the Image Quality Score and a weighted combination of the two evasion scores, with the participant-based evasion weighted at 0.7 and the organizer-based evasion at 0.3:

$$S_{\text{final}} = S_{\text{IQ}} \cdot (0.7 \cdot E_{\text{part}} + 0.3 \cdot E_{\text{org}}). \quad (5)$$

This formulation ensures that a submission ranks highly only when it succeeds on both fronts: a high-quality image that is trivially detected, or an undetectable image of poor quality, is penalized through the multiplicative coupling of the two factors.

4.2.2. Deepfake Detection Metrics

The deepfake detection sub-task is a binary classification problem, in which each system must decide whether a given frame is authentic or synthetically generated. Every score reported is an accuracy, i.e. the proportion of correct real-versus-deepfake decisions on a given subset, and all values lie in $[0, 1]$ with 1.0 denoting perfect classification. Rather than reporting a single accuracy over a homogeneous

Table 3

Evaluation subsets and weighting for the deepfake detection task.

Subset	Description	Weight
Baseline Deepfakes	Organizer-generated deepfakes	0.1
Organizers Real Data	Public authentic media	0.1
Organizers Real GT Data	Secret verified real subset	0.4
Participant Deepfakes (Weighted)	Realism-weighted participant deepfakes	0.4
Final Score	$S_{\text{final}} = 0.1 acc_{\text{base}} + 0.1 acc_{\text{real}} + 0.4 acc_{\text{GT}} + 0.4 acc_{\text{participants}}^w$	

test set, we decompose the evaluation into several complementary subsets, each probing a different aspect of detector robustness, and combine them into a weighted final score.

The first subset, *Baseline Deepfakes*, consists of deepfakes produced by the task organizers (Presented in Chapter 3) and measures accuracy against a controlled, known generation process; it contributes a weight of 0.1 to the final score. The *Organizers Real Data* subset contains authentic media released publicly in the starter kit and the generation task, and verifies that a system does not achieve high deepfake accuracy simply by over-predicting the fake class; it also carries a weight of 0.1. To guard against systems tuned to the public data, we additionally evaluate on *Organizers Real GT Data*, containing frames from the secret real set; its greater reliability as a clean reference is reflected in its higher weight of 0.4.

The remaining subsets assess generalization to unseen generators. *Participant Deepfake Data* comprises deepfake samples contributed by the other teams competing in the generation task. To ensure fairness, we filtered these contributions for similarity and retained only distinct generated subsets, so that a team submitting the same data multiple times is counted only once, even though it might have been present multiple times in the test dataset. The scored version of this subset, *Participant Deepfake Data (Weighted)*, reweights each submission by the realism score it obtained in the generation task, so that unrealistic generations which failed to meet the task constraints exert proportionally less influence on a detector’s measured accuracy; this weighted variant contributes 0.4 to the final score.

The **Final Score** for the modality, which serves as the primary ranking metric, is the weighted combination of these subset accuracies, with weights of 0.1, 0.1, 0.4, and 0.4 assigned to the Baseline Deepfakes, Organizers Real Data, Organizers Real GT Data, and weighted Participant Deepfake Data subsets respectively. The metrics are summarized in Table 3. This weighting deliberately emphasizes performance on the secret real ground-truth data and on realistic participant-generated deepfakes, rewarding detectors that generalize to high-quality, previously unseen content rather than overfitting to the publicly available material.

5. Results

5.1. Audio results

The results for the audio generation task are presented in Table 4. Team Go-To-Germany [29] achieved the highest overall score, obtaining the best balance between audio quality and evasion. Team Brihadaranyaka [30] and Team Chaboom [31] followed with very similar quality scores, but lower evasion values, indicating that the main difference between the strongest submissions was not only the perceptual quality of the generated speech, but also its ability to avoid detection.

A more detailed view of the audio generation results is provided in Table 5, which reports participant-side evasion, organizer-side evasion, and the raw metrics used to compute the quality component: NISQA, ECAPA-TDNN speaker similarity, WavLM speaker similarity, WER, and CER. Most competitive systems achieved low WER and CER values, showing that transcript preservation was generally strong across the submitted systems. Larger differences appeared in speaker similarity and evasion, especially between participant-side and organizer-side detectors. For example, Team Go-To-Germany and Team

Brihadaranyaka obtained almost identical quality scores, 0.7190 and 0.7159, respectively, but differed more clearly in evasion, with scores of 0.5985 and 0.4834.

The results for the audio detection task are shown in Table 6. Team Go-To-Germany obtained, once again, the best final score, with perfect accuracy on participant-generated deepfakes and strong performance across the remaining subsets. Team Deepfake1 [32] also achieved a high score, followed by Team Chaboom [31] and Team Twinkle [33]. The subset-level results show that, in the majority of the cases, the proposed systems performed well on authentic speech but struggled with participant-generated deepfakes.

Table 4
Deepfake Generation Results (Audio), ordered by Overall Score

Participant	Overall	Quality	Evasion
Team Go-To-Germany [29]	0.4304	0.7190	0.5985
Team Brihadaranyaka [30]	0.3461	0.7159	0.4834
Team Chaboom [31]	0.3378	0.7150	0.4725
leehyunwoo*	0.2695	0.5980	0.4507
ryu-tae-ung*	0.2594	0.6528	0.3974
Team Deepfake1 [32]	0.2224	0.6738	0.3301
gwangseop_lee*, leehowoo*, kim_minseo*	0.2151	0.6793	0.3166
k20221289 [34]	0.2047	0.6635	0.3086
Team Twinkle [33]	0.1951	0.6969	0.2800
Team CS Morgan Lab [35]	0.1806	0.4931	0.3662

* Participants marked with an asterisk did not submit a corresponding working notes paper.

Table 5
Audio Deepfake Generation Detailed Metrics, ordered by Overall Score, The table presents the Evasion scores, against participant models (Eva. (P)), against organizer models (Eva. (O)), and the five raw metrics calculated for each submission.

Participant	Overall	Eva. (P)	Eva. (O)	NISQA	ECAPA	WavLM	WER	CER
Team Go-To-Germany [29]	0.4304	0.6142	0.5618	2.9729	0.6148	0.9341	0.0499	0.0207
Team Brihadaranyaka [30]	0.3461	0.5278	0.3799	3.0976	0.5667	0.9398	0.0783	0.0539
Team Chaboom [31]	0.3378	0.5521	0.2868	3.1405	0.5959	0.8622	0.0793	0.0393
leehyunwoo*	0.2695	0.5308	0.2639	2.7576	0.2885	0.8257	0.1218	0.1775
ryu-tae-ung*	0.2594	0.5129	0.1278	3.0413	0.3214	0.8294	0.0990	0.0610
Team Deepfake1 [32]	0.2224	0.4343	0.0868	3.0876	0.4109	0.8375	0.0932	0.0552
gwangseop_lee*, leehowoo*, kim_minseo*	0.2151	0.4303	0.0514	3.0467	0.4360	0.8417	0.0730	0.0408
k20221289 [34]	0.2047	0.4090	0.0743	3.1131	0.4460	0.8450	0.1550	0.1217
Team Twinkle [33]	0.1951	0.3868	0.0306	3.1787	0.4499	0.8393	0.0642	0.0315
Team CS Morgan Lab [35]	0.1806	0.5217	0.0035	1.5129	0.3657	0.8063	0.1394	0.0872

* Participants marked with an asterisk did not submit a corresponding working notes paper.

5.2. Audio presented methods

5.2.1. Sub-task 1: Deepfake generation

Team Brihadaranyaka’s method [30] uses an ensemble of three models zero-shot voice cloning models: F5-TTS [38], IndexTTS [4] and XTTS v2 [2]. Voice cloning is employed with these methods and the best result is selected according to the speaker embedding cosine similarity between the generated audio and the voice reference. The selected audio is then post-processed for evasion using MP3 codec pass-through,

Table 6

Audio Deepfake Detection Detailed Metrics, ordered by Final Score. The table presents the accuracy metric on each subset of data: Organizer-generated deepfakes (Org. DF), Real audio files from the generation task (Org. Real), Real audio files from the Ground Truth/Secret dataset (Org. Real GT, Participant-generated deepfakes (Part. DF), and Participant-generated deepfakes, weighted by their overall quality score. (Part. DF(W)).

Participant	Score	Org. DF	Org. Real	Org. Real GT	Part. DF	Part. DF (W)
Team Go-To-Germany [29]	0.9522	0.9719	1.0000	0.8875	1.0000	1.0000
Team Deepfake1 [32]	0.8814	0.8307	0.9521	0.8917	0.8712	0.8661
Team Chaboom [31]	0.7564	0.5411	0.8750	0.8438	0.6950	0.6932
gwangseop_lee*	0.7267	0.3698	0.9500	0.9625	0.5295	0.5243
Team Twinkle [33]	0.7017	0.2427	1.0000	0.9979	0.4343	0.4457
leehowoo*	0.6801	0.3646	0.7958	0.8542	0.5579	0.5559
Team CS Morgan Lab [35]	0.5918	0.1646	0.9896	0.9938	0.2008	0.1972
ryu-tae-ung*	0.5769	0.1057	1.0000	1.0000	0.1607	0.1658
botter [36]	0.5595	0.4182	0.5396	0.6271	0.5377	0.5323
cc_rmws [37]	0.5551	0.2979	0.8146	0.8625	0.2407	0.2472
k20221289 [34]	0.5292	0.1459	0.9333	0.8854	0.1707	0.1678
haripriya, petchiammal	0.5253	0.8349	0.1833	0.1437	0.9186	0.9151
Team Brihadaranyaka [30]	0.5092	0.3594	0.5813	0.6083	0.4336	0.4295
minbro_01*	0.4883	0.5110	0.6208	0.5917	0.3425	0.3461
kim_minseo*	0.4405	0.1953	0.6396	0.6229	0.2691	0.2697

* Participants marked with an asterisk did not submit a corresponding working notes paper.

synthetic room impulse response convolution, and calibrated micro-noise injection targeting 48 dB SNR.

Team Team CS Morgan Lab [35] solution is based on XTTS-v2 [2], a multilingual zero-shot voice-cloning model. The reference is chosen using a short/medium/long utterance-category heuristic: same-category references of at least 3 seconds are preferred, otherwise the longest available reference is used. To improve evasion against anti-spoofing systems, the generated waveform is passed through a four-stage channel-domain post-processing chain designed to mimic realistic recording conditions: room impulse response convolution with $RT60 = 0.3$ s mixed at 15%, Gaussian noise at 38 dB SNR, a 4th-order Butterworth low-pass filter at 7.5 kHz blended at 90/10 to simulate codec roll-off, and a $\pm 2\%$ time-stretch perturbation. The final output is saved as a 16 kHz PCM-16 waveform.

Team Twinkle [33] uses XTTSv2 [2] to generate the audios and uses a strategy to maximize the reference audio length used for conditioning the speech language model by merging all available speech recordings from the same speaker. A notable contribution is that, for conditioning the vocoder component, they use an optimized speaker embedding by minimizing the distance to the 30 audio reference embeddings from that speaker and maximizing the distance from other speaker’s embedding recordings.

Team Deepfake1[32] have proposed a method based on XTTS-v2 [2] with a fallback to SpeechT5 [39] combined with HiFi-GAN [40] and SpeechBrain x-vectors [41]. To make their audio deepfakes more credible, they simulated a Mu-law codec, combined with a low-pass filter and dithering. In doing so, the high-frequency spectral fingerprint of the used TTS systems is removed from the generated speech. To ensure the same output volume, the audio is further processed using an amplitude-based super-reference.

Team Chaboom [31] employed an F5-TTS [38] for generating the audio deepfakes. Their pipeline consisted of four different steps, each with a well-defined scope. Firstly, they did an analysis to determine the best length for the reference selection, then used the F5-TTS for inference, followed by a post-processing pipeline, and finally applied a fade-in/fade-out effect. The post-processing pipeline consisted of high-frequency removal using a Butterworth filter with a cut-off frequency $f_c = 7500$ Hz, a silence trimmer for both ends of the audio, and a LUFS Loudness normalization module.

Team Go-To-Germany [29] developed a hybrid TTS system combining GLM-TTS [42], F5-TTS [38], XTTS v2 [2], and CosyVoice 3 [3], assigning different speakers to different models to increase generator diversity and reduce detectability. The authors employed six lightweight internal detectors to evaluate generated speech and used speaker-level cherry-picking together with temperature tuning to balance speech quality and evasion capability. Their analysis showed that combining multiple TTS architectures improved detection evasion while preserving speech quality. The work also highlighted the trade-off between audio quality and detectability and demonstrated how insights obtained during generation could be transferred to improve deepfake detection systems. This team achieved the best results for the audio deepfake generation task.

5.2.2. Sub-task 2: Deepfake detection

Team cc_rmws [37] proposed a system which incorporates a multi-view feature fusion approach, extracting three complementary representations in parallel: WavLM-large [11] self-supervised features for high-level semantic and contextual information, Linear Frequency Cepstral Coefficients (LFCC) for spectral structure, and Modified Group Delay (MGD) features for subtle phase discontinuities that neural vocoders tend to introduce. These are combined through a Triple-Fusion module and fed into a lightweight Dual-Branch back-end, where one branch models long-range speech consistency, while the other focuses on short-term local acoustic anomalies. The model was trained on the ASvspoof 2019 LA [43] dataset as a proxy, achieving strong accuracy on real speech subsets but struggling on participant-generated spoofed speech – a gap the authors attribute to the distribution mismatch between the proxy training data and the unseen generation methods present in the challenge’s hidden test set.

Team Brihadaranyaka [30] uses an ensemble of three public audio detectors: MelodyMachine [44], Heem2 [45], and Gustking [46]. In the first variant, the detector scores are rank-normalized and averaged, while in the second variant the three detector outputs are directly averaged.

Team Team CS Morgan Lab [35] fine-tunes Wav2Vec2-XLS-R-300M [47] that was originally pretrained for audio spoof classification. The detector is partly fine-tuned: The convolutional frontend and the first 12 of 24 transformer layers are frozen, leaving only approximately half of the parameters trainable. Training uses label-smoothed cross-entropy, AdamW, and a cosine learning-rate schedule. The training corpus that was used by the team is Deepfake Audio Detection Dataset (v4) [48], containing 1,866 class-balanced clips, split into 1,586 training and 280 validation examples. During inference, the system estimates one prediction for four random 5 second windows, and averages the fake probabilities.

Team Twinkle [33] uses a system based on AASIST spoofing detector [49], which is fine-tuned using a self-augmented training strategy, extending public audio deepfake detection datasets with fake audios generated by their own system that was developed for the generation task.

Team Deepfake1 [32] approach the audio detection task by combining two scoring functions that share a common K-NN trained on self-supervised embeddings. The two branches used for scoring use ensemble weights from WavLM [11] and wav2vec2 (base) [46] with 0.6 and 0.4, respectively. The pooling cosine-anomaly score from these two different encoders is used to compute the distance to the real distribution, which is then thresholded to determine the real or fake label.

Team Chaboom [31] used a single-model pipeline approach for audio deepfake detection. The model used is AASIST [49], using a pre-processing technique to ensure the required audio specification for this model, as well as a chunk-based prediction, which are then aggregated w.r.t. the mean (0.8) and maximum (0.2) fake probabilities before applying a fixed threshold (0.5) to determine the belonging class.

Team Go-To-Germany [29] proposed a multi-backbone self-supervised learning (SSL) ensemble for audio deepfake detection. Their approach combines frozen WavLM-Large [11], Wav2Vec2-XLS-R-300M [46], ECAPA-TDNN [10], and x-vector [41] feature extractors with independent logistic regression classifiers whose outputs are fused through a weighted ensemble. To improve robustness against unseen deepfake generators, the method exploits the complementary representations learned by different SSL backbones. The submitted system achieved a final score of 0.9522, including perfect detection of participant-generated deepfakes, achieving the best results for the audio deepfake detection task.

5.3. Images results

The results for the **Generation** task are presented in Table 7. Two teams achieved markedly stronger results than the rest: Team Brihadaranyaka and Team Go-To-Germany. While several other teams attained high evasion scores, only these two managed to satisfy both the identity and landmark constraints. A visual comparison of their approaches is shown in Figure 4, in the columns corresponding to the best quality scores. Both teams preserved the facial proportions and positions defined by the landmarks, keeping them closely aligned with the target. The two teams differed considerably in their treatment of the frames: Team Brihadaranyaka favored a more realistic approach that resembles the original videos, whereas Team Go-To-Germany employed models that diversified the background and clothing while keeping the identities largely constant, yielding high-quality images.

The evasion scores vary considerably across submissions, though higher-quality images generally also achieve high evasion scores. Even so, the participant evasion scores for all but four submissions fall below 0.5, meaning the images were classified as deepfakes more often than not by other participants. The top three submissions achieved evasion scores above 0.5. Evading the organizer models, on the other hand, was generally straightforward for the majority of submissions. Because the organizer models were open-source, these results suggest that deepfake detectors are typically not hard to fool, particularly when the generation mode differs from the training data (many open-source models are trained on identity swapping, for example).

Most generated images attained a relatively high identity similarity score. Landmark distance scores, by contrast, varied widely, with roughly half of the submissions achieving weak scores in this column. This may reflect the difficulty of generating images while simultaneously satisfying the imposed constraints.

The results for the **Detection** task are presented in Table 8. Most teams hover just above a 0.5 final score, indicating that their detections were close to random in most cases. The best approaches reach a final weighted accuracy of 0.7, demonstrating strong classification ability, though not strong enough for real-world applications.

A first thing to notice is the difference between real and deepfake accuracies for some models. Near-perfect real scores combined with near-zero deepfake scores suggest improper calibration of the decision threshold, which may stem from differences between the training and evaluation datasets. Participants were free to train on any data they chose, so these differences are not surprising. The best models provided a balanced approach, outputting similar accuracy scores for real and deepfake data.

The weighted accuracy values on the participant data was smaller than the non-weighted counterpart, proving that high-quality generated images are harder to detect.

5.4. Images presented methods

5.4.1. Sub-task 1: Deepfake generation

Team MarsadLab [51] approached the generation task with a warp-based generation technique, transforming real video frames to match target landmarks. Their pipeline extracts 20 frames per development video, selects the closest one to the required facial landmarks, minimizing deformation needed. Identity embeddings are extracted with InsightFace (buffalo_s) [54], and the selected face is aligned to the target geometry using a PiecewiseAffineTransform over roughly 200 key facial landmarks plus boundary anchors. The warped output is refined with GFPGAN [55] blind face restoration and then passed through lightweight anti-forensic post-processing (JPEG round-trip compression, additive Gaussian sensor noise, and subtle 0.3-pixel Gaussian blur) to suppress artifacts exploited by detectors. For the generation task, the team asked for a late-submission option where their generated images would not be included in the Detection task. Therefore, the Evasion score for participants is set at 0.

Team Deepfake1 [32] used a five-stage hybrid pipeline, built on Stable Diffusion 1.5 that decouples identity conditioning from landmark conditioning. A top 5 frames from 30 equidistant frames is selected, based on a composite sharpness/frontality/area score. Identity is extracted using InsightFace from those faces and averaged to obtain an identity vector. The target landmarks are rendered into a region-coded

Table 7

Image Deepfake Generation Results, ordered by Overall Score. The Overall score is a product of Quality and Evasion. The Evasion score is split into Evasion of other Participant models (Evasion Part.) and evasion of organizers models (Evasion org.). The values reported for Landmark distance and Identity Similarity are the raw values, before the scalings presented in Chapter 3.

Participant	Overall	Quality	Evasion	Evasion Part.	Evasion Org.	Landmark Distance	Identity Similarity
Team Brihadaranyaka [30]	0.4364	0.7056	0.6185	0.5603	0.7541	0.0154	0.6239
Team Go-To-Germany [50]	0.4170	0.6188	0.6739	0.5762	0.9019	0.0065	0.5366
Team Chaboom [31]	0.2486	0.3823	0.6503	0.5657	0.8478	0.0076	0.1749
ryu-tae-ung*	0.2298	0.4672	0.4918	0.3885	0.7330	0.1771	0.6803
gwangseop_lee*, kim_minseo*, leehowoo*	0.2264	0.4620	0.4902	0.3822	0.7420	0.1197	0.6575
Team Deepfake1 [32]	0.1541	0.4000	0.3854	0.3574	0.4505	0.0074	0.4097
Team Twinkle [33]	0.1471	0.2552	0.5767	0.5269	0.6927	0.1305	0.5458
k20221289 [34]	0.1225	0.3705	0.3306	0.3185	0.3589	0.0092	0.3100
Team MarsadLab [51]	0.0967	0.5631	0.1717	0.0000	0.5722	0.0093	0.5160
Team CS Morgan Lab [35]	0.0000	0.0000	0.4167	0.3350	0.6073	0.3802	0.0000
leehyunwoo*	0.0000	0.0000	0.3018	0.2373	0.4523	0.2580	0.0000

* Participants marked with an asterisk did not submit a corresponding working notes paper.

FaceMesh [56] conditioning image, with each semantic facial region (face oval, eyes, eyebrows, lips) drawn in a distinct RGB color so the ControlNet [57] attends to labeled structure rather than a generic outline. Generation combines a MediaPipe-FaceMesh ControlNet (structure/pose) with IP-Adapter-FaceID [58] (identity, injected via cross-attention), using DDIM [59]. A Soft Landmark Alignment stage then corrects residual landmark error with a clamped 4-Degree Of Freedom (DOF) Similarity Transform blended through a feathered face mask, avoiding the warping artifacts that frequency detectors exploit. A final ensemble PGD [60] step attacks two ViT-based public detectors, nudging detector confidence toward the real class.

Team Brihadaranyaka [30] treats Generation and Detection as one unified system, deliberately reusing a single proxy detector, trained on FaceForensics++, Celeb-DF and another 140k real and fake faces, as both a PGD attack target in Generation and an ensemble member in Detection. The paper describes the three submitted runs: The first run uses affine transforms to warp the closest frame in the closest to the target, then the image is attacked with PGD adversarial perturbation targeting the targeting the proxy detector and a light Gaussian blur. The second and third run InstantID [61] on Stable Diffusion XL [8], conditioning on a 512-dimensional InsightFaceidentity embedding via decoupled cross-attention and a 5-keypoint reduction ControlNet. Submission 2 introduces noise injection, while submission 3 drops it for blur and JPEG compression.

Team k20221289 [34] uses Stable Diffusion 1.5 conditioned by both ControlNet and IP-Adapter to synthesize identity-preserving faces matching the target 478-point MediaPipe landmarks. During development they addressed three failure modes by tuning generation parameters: identity drift was reduced by raising the IP-Adapter scale from 0.6 to 0.8, structural landmark misalignment was corrected by increasing the ControlNet conditioning scale from 1.0 to 1.2, and over-sharpened outputs were softened by lowering the classifier-free guidance scale from 7.5 to 5.0.

Team Chaboom [31] uses a four stage diffusion pipeline with landmark constrains. A frontal source frame was selected per subject as the identity reference, along with extracted person attributes extracted with CLIP (ViT-B/32) [62] for generation prompt construction. Landmarks were rendered onto a canvas image using different colors for different landmark regions. The pipeline integrates three components into a Stable Diffusion 1.5 pipeline: ControlNet for geometry, IP-Adapter for identity, and Realistic Vision V6.0 [63] as a base U-Net [64] checkpoint, finetuned for realistic human portraits. The pipeline was validated using a subset of Celeb-DF v2.

Table 8

Image Deepfake Detection Results, ordered by Final Score. The table presents the performance of the participants' models against: Organizer-Generated Deepfakes (Org. DF), Real data distributed by the organizers (Org. Real), Real data from the Ground Truth/Secret Dataset (Org. Real GT), Deepfakes Generated by participants (Part. DF), and Deepfakes Generated by participants, weighted by the quality scores (Part. DF (W)).

Participant	Score	Org. DF	Org. Real	Org. Real GT	Part. DF	Part. DF (W)
minbro_01*	0.7103	0.7995	0.6047	0.6607	0.7933	0.7641
Team Go-To-Germany [50]	0.6987	0.9939	0.4324	0.5729	0.8822	0.8171
k20221289 [34]	0.6936	0.6660	0.9056	0.7335	0.6774	0.6077
isurukalhara*	0.6297	0.1488	0.9409	0.9541	0.4611	0.3477
Team Deepfake1 [32]	0.6253	0.6919	0.8426	0.9102	0.3535	0.2694
botter [36]	0.6038	0.7928	0.6241	0.4301	0.7476	0.7252
nataliemorad*	0.5964	0.6284	0.7129	0.6557	0.6217	0.5000
cc_rmws [37]	0.5858	0.5266	0.6491	0.5948	0.6363	0.5758
phantom [52]	0.5772	0.5710	0.3426	0.3204	0.9009	0.8942
Team MarsadLab [51]	0.5766	0.7382	0.6026	0.5040	0.6540	0.6022
cyril_xu	0.5648	0.3858	0.8103	0.7735	0.4546	0.3396
jayasree [53]	0.5562	0.6454	0.5309	0.6287	0.5557	0.4677
haripriya, petchiammal	0.5448	0.3179	0.8129	0.8353	0.2778	0.2439
ryu-tae-ung*	0.5402	0.6859	0.4203	0.3473	0.6905	0.7267
leehowoo*	0.5362	0.5343	0.4756	0.5609	0.5883	0.5271
Team Brihadaranyaka [30]	0.5358	0.5580	0.5388	0.6896	0.4208	0.3756
Team CS Morgan Lab [35]	0.5197	0.0950	0.8371	0.9641	0.1576	0.1021
Team Twinkle [33]	0.5096	0.7587	0.3897	0.2265	0.7028	0.7604
gwangseop_lee*	0.5092	0.8968	0.1012	0.1727	0.8404	0.8509
kim_minseo*	0.5087	0.8973	0.0997	0.1717	0.8404	0.8507
mohamedbasem*	0.5052	0.0244	0.9988	1.0000	0.0574	0.0072
Team Chaboom [31]	0.5022	0.8351	0.1826	0.2874	0.7098	0.7136

* Participants marked with an asterisk did not submit a corresponding working notes paper.

Team CS_Morgan Lab [35] used a few-shot identity-preserving synthesis from one reference video per identity. The extracted landmarks were passed to a Stable Diffusion v1.5 pipeline combining ControlNet for landmark adherence and IP-Adapter-FaceID for identity, sampled with UniPC. The post-processing included a momentum-PGD attack against an ensemble of four models (Xception, EfficientNet-B4/B7, ViT-B/16), JPEG compression and blurring for evasion.

Team Go-To-Germany [50] built their pipeline from three off-the-shelf components: FLUX.1-dev [65] generated 512×512 faces from a fixed prompt, ControlNet-Union-Pro-2.0 [66] enforced facial geometry (the supplied MediaPipe landmarks were rendered into an OpenPose-style control image to match the ControlNet's pretraining convention), and PuLID [67] injected identity from an embedding averaged over up to 20 diversity-sampled reference crops. Per identity, 11 candidates were generated and filtered by landmark distance and identity similarity, then downsampled to the 256×256 submission resolution. Their main contribution lay in the adversarial post-processing: a custom PGD attack ($\epsilon = 2/255$, 25 iterations) jointly optimized against 12 white-box detectors. A differentiable JPEG simulation at quality $Q = 80$ was applied to the perturbed image before each forward pass, so that gradients reflected the post-compression image actually seen by the detectors, while Input Diversity (DI) and Expectation over Transformation (EoT) [68] further improved transferability to unseen detectors. Adaptive per-image loss weighting and a two-stage warm-start attack were added to handle detectors with stubborn failure cases.

Team Twinkle [33] approached image generation as reference-driven face reenactment rather than synthesis from scratch. For each identity, development videos were decoded and a reference frame was selected. For each of the 1,002 constraint entries, the driving frame was chosen as the video frame minimizing mean Euclidean landmark distance to the target 478-point pose. LivePortrait [69] was used

to transfer the driving pose and expression onto the identity reference through its implicit 3D keypoint motion representation. The main contribution of the paper is a three-stage post-processing pipeline that corrected the raw reenactment’s characteristic failure modes: Reinhard-style LAB colour transfer to fix photometric drift [70], CLAHE [71] plus unsharp masking to recover decoder-induced blur, and a RANSAC-estimated [72] similarity transform that re-aligned the output to the target landmarks.

5.4.2. Sub-task 2: Deepfake detection

Team MarsadLab [51] uses a two-branch weighted ensemble for detection. The primary branch is a CLIP-ViT-L/14 detector following Yermakov et al. [73], where only the Layer Normalization parameters (0.03% of the model) are fine-tuned to preserve CLIP’s general representations while adapting to deepfake detection; faces are detected and cropped with MTCNN [74] (with a center-crop fallback) before classification. The secondary branch is an ImageNet-pretrained EfficientNet-B4 CNN classifier that contributes fine-grained texture and artifact cues. The two fake-class probabilities are combined as a weighted sum, favoring the CLIP branch with a 0.7 weight. They also use a threshold of 0.5 for prediction.

Team Deepfake1 [32] used a four model ensemble for deepfake detection. The first branch is a CLIP ViT-L/14 feature k-NN: reference (real) frames are embedded and indexed, and a test frame’s mean cosine distance to its five nearest real neighbors is turned into an anomaly score against a mean+2-std threshold. The second is CLIP zero-shot classification, using averaged real and fake text-prompt prototypes and taking the fake-class softmax probability. The third model is an open-source ViT deepfake classifier with horizontal-flip test-time augmentation, trained on a Kaggle dataset. The fourth method is an FFT frequency-ratio detector that compares mean log-magnitude in high- versus low-frequency annular spectrum regions, min-max normalized over a calibration set.

Team Brihadaranyaka [30] uses a pool of public detectors (ViT, Swin [75], SigLIP [76], and EfficientNet backbones), with two approaches: (1) normalizing the 8 detector scores and averaging, and the main contribution, (2) dropping the lowest-performing detectors, polarity-correcting the remaining ones, with their scores feeding a logistic-regression model trained on a set of 394-sample calibration set created by the authors. The most uncertain predictions are re-scored by an LLM (Claude Sonnet 4.6) and blended back in, though the team reports this adds no measurable gain.

Team botter [36] proposes a method involving addressing the face-scale mismatch between the aligned training data (DF40 [77]) and the full-frame CLEF images through a three-tier face-alignment cropper. Each aligned crop is fed to three different base detectors: a DF-40 finetuned CLIP classifier, GenD (the same CLIP encoder adapted by tuning only Layer Normalization parameters with feature normalization and metric-learning objectives), and NPR (an artifact detector operating on neighboring-pixel-relation residuals classified by a truncated ResNet-50). The prediction is done by a weighted sum of the probabilities.

Team k20221289 [34] proposes a three-component ensemble motivated by distinct fake categories. A fine-tuned CLIP ViT-L/14 encodes 224×224 face crops into 768-dimensional features and targets GAN and FaceSwap artifacts through its language-aligned semantic feature space. A fine-tuned DINOv2 ViT-L/14 extracts the 1024-dimensional CLS token and, thanks to its self-supervised visual pretraining, is emphasized toward diffusion-generated images and their texture/structural irregularities. A third signal comes from InsightFace, which acts as a face-quality proxy. The three fake probabilities are standardized and fused by a logistic-regression stacking meta-learner. The CLIP and DINOv2 [78] detectors were trained on a mix of Dev Video real crops, FaceForensics++ real and GAN/FaceSwap fake faces, and Stable Diffusion text-to-image outputs.

Team Chaboom [31] implemented an inference-only pipeline using three open-source pretrained detectors combined via a weighted average: a primary deepfake detector with the largest weight, an auxiliary deepfake detector with a smaller weight, and an AI-generated image detector with the smallest weight. Each detector produced four scores corresponding to four preprocessing views: original, flip, face crop, and wide face crop. For each detector, these view scores were combined using a mix of a maximum score and a weighted average, before the three detectors were fused into the final ensemble

score.

Team jayasree [53] proposes a dual-stream spatial-frequency framework for generalized deepfake detection. The method runs two parallel EfficientNet-B0 backbones – one processing standard RGB images to capture spatial cues like texture inconsistencies and blending artifacts, and another processing Fast Fourier Transform (FFT)-based frequency representations to detect hidden spectral irregularities introduced during image synthesis. A key innovation is a consistency regularization term added to the training loss, which encourages the two branches to produce aligned feature representations by exploiting the observation that real images tend to have coherent spatial and frequency characteristics while manipulated ones do not. The model was trained on multiple manipulation categories from the FaceForensics++ dataset with heavy augmentation to promote generalization, achieving a final score of 0.5562 on the challenge’s unseen test set.

Team cc_rmws [37] proposes a HF-Swin detection framework built on a Swin-Small backbone that jointly models RGB semantic features and high-frequency residual information. Rather than training on fixed deepfake datasets, the system employs a Self-Blended Images (SBI) [79]-based online pseudo-forgery generation strategy, dynamically constructing forged training samples from real face frames extracted from FaceForensics++ by applying random color, geometric, texture, and frequency perturbations. The high-frequency residual is computed as the difference between the input image and its average-pooled version, highlighting local texture anomalies and edge discontinuities. A boundary-guided module further applies spatially-aware reweighting to this residual before it enters the high-frequency branch, directing the model’s attention toward facial boundary regions and blending transition areas rather than irrelevant background noise. The two feature streams are concatenated and passed to a classification head, achieving a final score of 0.5858 on the hidden test set.

Team CS_Morgan Lab [35] used an ensemble system composed of two fine-tuned on CelebDF v2 (a ViT-base and SigLIP2 detector), each with 4× test-time augmentation and combined via equal-weight averaging. Uncertain samples were routed to a frozen Qwen2.5-VL-3B judge with chain-of-evidence prompting. The submission threshold was fixed at 0.5.

Team Go-To-Germany [50] used a two-detector max-probability ensemble of off-the-shelf specialists: SigLIP+DINOv2 for full-image AI-generated content and GenD-DINOv3 [80] for face manipulation. Inputs were first matched against perceptual hashes of the team’s own 1,002 generations. Model choice was driven by a self-curated 10,000-image benchmark (FaceForensics++ and BitMind [81]), where GenD-DINOv3 dominated face manipulations and SigLIP+DINOv2 dominated AI-generated content, motivating the complementary pairing. The min-probability threshold was calibrated on a separate 7,000-image set by count-based net gain – a criterion the authors later identify as the flawed step behind their weak real-image accuracy.

Team Twinkle [33] adopted the Cross-EfficientViT architecture presented in [54], which couples an EfficientNet-B0 spatial feature extractor with a Vision Transformer cross-attention module so the classifier can attend selectively to spatially localized manipulation artifacts in each cropped face. They compared two variants on a held-out image test set: a standalone EfficientNet-B0 baseline and the full Cross-EfficientViT, and submitted the latter, which performed best internally. Classification used a fixed decision threshold of 0.5 without validation-set calibration, which the authors identify as the decisive weakness.

Team phantom [52] used an ensemble of three different pretrained models with diverse architectures: a Swin Transformer, an EfficientNet-B3 CNN, and a CLIP ViT-B/32 vision encoder fully fine-tuned with a custom classification head. Each was trained independently on a combined corpus of roughly 1.04 million images assembled from several public real-face and deepfake datasets (FFHQ [82], the 140k Real and Fake Faces set, DFDC [83] face crops, and Stable-Diffusion face images), supplemented with additional diffusion-based AI-generated faces to cover recent generators. Each image was scored with multi-scale test-time augmentation (six views combining three scales with horizontal flips, logits averaged per model).

6. Discussion

6.1. Audio

A wide range of strategies was deployed for the audio modality, both for generation and detection. In the generation subtask, most teams relied on modern zero-shot or few-shot text-to-speech systems, with XTTS v2 and F5-TTS appearing most frequently among the submitted approaches. Several participants also explored multi-model generation pipelines, combining systems such as F5-TTS, IndexTTS, XTTS v2, CosyVoice, GLM-TTS, SpeechT5, or HiFi-GAN-based backends. The most successful systems did not treat voice cloning as a single-model inference problem, but rather as a selection and optimization problem, where different synthesis backbones, reference utterances, sampling settings, and post-processing configurations were compared before submission.

The best audio generation score was obtained by Team Go-To-Germany [50], whose submission achieved an overall score of 0.4304, with a quality score of 0.7190 and an evasion score of 0.5985. Their approach combined multiple TTS backbones and selected outputs in a detector-aware manner, using internal lightweight detectors and speaker-level tuning to balance intelligibility, speaker preservation, and evasion. Team Brihadaranyaka [30] and Team Chaboom [31] followed closely, with final scores of 0.3461 and 0.3378, respectively. Both systems also illustrate an important trend of the task: high generation scores required not only good speech quality, but also an explicit strategy for reducing detectability. In contrast, submissions with relatively high audio quality but lower evasion were penalized by the multiplicative scoring formula, confirming that perceptual realism alone is insufficient for this benchmark.

A consistent observation across generation submissions is that transcript preservation was generally easier to achieve than robust detector evasion. Several systems obtained low WER and CER values, indicating that current TTS models can reproduce the target sentence reliably in the provided few-shot setting. For example, the top-ranked submission achieved WER = 0.0499 and CER = 0.0207, while other competitive systems also obtained WER values below 0.10. This suggests that content fidelity is no longer the dominant bottleneck for audio deepfake generation under these conditions. Instead, the main limitations were speaker-identity preservation across different embedding models, perceptual naturalness after post-processing, and evasion against strong anti-spoofing detectors.

Speaker similarity results reveal an important methodological issue. WavLM-based similarity scores were consistently higher than ECAPA-TDNN scores for most submissions. The top submissions obtained WavLM similarities above 0.86, and several exceeded 0.93, whereas ECAPA-TDNN values were substantially lower. This gap indicates that speaker preservation depends strongly on the embedding model used for evaluation. WavLM appears more tolerant to broad speaker characteristics, while ECAPA-TDNN is more sensitive to fine-grained speaker-discriminative cues. The use of both encoders was therefore beneficial, since it reduced the risk that systems optimized only for one speaker-representation space would be overestimated.

Post-processing was one of the most common generation-side strategies. Participants applied MP3 codec pass-through, room impulse response convolution, Gaussian noise, low-pass filtering, Butterworth filtering, time stretching, Mu-law companding, dithering, loudness normalization, silence trimming, and fade-in/fade-out operations. These transformations were intended either to simulate realistic recording conditions or to suppress spectral fingerprints introduced by TTS systems and vocoders. However, the results show that this strategy has a narrow operating range. Light and controlled post-processing sometimes improved evasion without destroying intelligibility, but stronger perturbation pipelines often reduced perceptual quality or failed to transfer to organizer-side detectors. The lowest organizer-side evasion values show that simple channel simulation is not sufficient against robust spoofing systems.

The difference between participant-side and organizer-side evasion further highlights this limitation. For almost all systems, evasion against participant detectors was higher than evasion against organizer detectors. The most extreme cases obtained reasonable participant evasion but almost no organizer evasion, showing that generated audio could exploit weaknesses of some submitted detectors while still being easily recognized by stronger or better-calibrated countermeasures. This suggests that audio

deepfake evasion remains fragile when optimized against a limited or mismatched set of surrogate detectors. Future generation systems will likely need stronger adversarial objectives, better vocoder diversity, and more realistic channel modeling to generalize across detector families.

For the detection subtask, the strongest results were achieved by systems that combined complementary speech representations rather than relying on a single handcrafted feature or a single anti-spoofing backbone. Team Go-To-Germany [29] obtained the highest final score, 0.9522, with 1.0000 accuracy on participant-generated deepfakes and 0.8875 accuracy on the hidden real ground-truth subset. Their system fused WavLM-Large, Wav2Vec2-XLS-R, ECAPA-TDNN, and x-vector representations, showing that multi-backbone self-supervised and speaker-recognition features can provide strong robustness to heterogeneous synthetic speech. The second-ranked system, submitted by Team Deepfake1 [32], also used self-supervised speech embeddings, combining WavLM and wav2vec2 with a K-NN anomaly-style decision rule over the real distribution, and achieved a final score of 0.8814.

The most frequently used detection components were WavLM, wav2vec2 or XLS-R representations, AASIST-based anti-spoofing models, LFCC features, ECAPA or x-vector speaker embeddings, and score-level ensemble fusion. Some teams also used Modified Group Delay features, mel or spectral descriptors, chunk-based inference, random-window averaging, or K-NN scoring against authentic speech embeddings. This diversity reflects the fact that audio deepfake detection benefits from combining complementary cues: high-level phonetic and contextual consistency, low-level spectral artifacts, phase irregularities, vocoder fingerprints, and speaker-representation anomalies.

Several approaches worked well on real speech but failed on participant-generated deepfakes. This was particularly visible for systems trained mainly on public proxy datasets such as ASVspoof, WaveFake, or other external anti-spoofing corpora. These models often achieved high accuracy on organizer real data and hidden real recordings, but much lower accuracy on participant deepfakes. For instance, some submissions exceeded 0.93 accuracy on organizer real data while scoring below 0.25 on participant-generated fake speech. This pattern indicates a strong domain mismatch between standard anti-spoofing training data and the newer zero-shot voice cloning systems used by the challenge participants. It also confirms that high validation performance on external corpora does not necessarily translate to robustness on unseen generators.

In contrast, methods that incorporated model diversity, task-specific self-augmentation, or real-distribution scoring generalized better. The AASIST-based submission from Team Twinkle [33], for example, benefited from adding self-generated clips to the training data, while the best-performing SSL ensemble benefited from the fact that different backbones captured different artifact families. These results support the view that audio deepfake detection is not solved by a single representation. A detector trained to recognize one generator family may fail when the synthesis model, vocoder, speaker set, sampling conditions, or post-processing chain changes. Robust systems should therefore combine heterogeneous features and be calibrated on data that resemble the intended deployment distribution.

Another important finding is the role of threshold calibration. Several participant papers reported strong internal validation performance but weaker official results due to biased decision thresholds or class-specific failure modes. Some detectors were conservative and classified too many files as real, leading to low fake detection accuracy. Others were overly aggressive and misclassified authentic recordings as deepfakes. Because the official score assigns large weights to both hidden real recordings and participant-generated deepfakes, this type of imbalance can strongly reduce the final score. The category-level results, therefore, show that calibration is not a secondary implementation detail, but a central requirement for robust deepfake detection.

Overall, the audio results show that the detection task was more successfully addressed than the generation task. The best detection systems achieved high scores and strong generalization to participant-generated deepfakes, whereas the best generation systems remained below 0.45 overall due to the combined difficulty of preserving speech quality, speaker identity, and detector evasion. This asymmetry suggests that, in the current benchmark setting, defenders benefited from ensemble diversity more than generators benefited from TTS diversity and post-processing. At the same time, the existence of generation submissions with strong intelligibility, high WavLM similarity, and non-trivial evasion confirms that modern voice cloning systems are already capable of producing challenging synthetic

speech from limited reference material.

6.2. Images

6.2.1. Generation

The systems submitted to the Generation sub-task spanned a broad methodological spectrum, from purely geometric manipulation to full generative synthesis. At one end, several teams framed the problem as warping or reenactment: Team MarsadLab aligned real frames to the target landmarks through a PiecewiseAffineTransform, Team Twinkle performed reference-driven face reenactment using LivePortrait’s implicit 3D keypoint motion representation, and one of Team Brihadaranyaka’s runs relied on a direct affine warp of the closest matching frame. At the other end, the majority of participants adopted diffusion-based synthesis, with generative backbones including Stable Diffusion 1.5 (Teams Deepfake1, k20221289, Chaboom, and CS_Morgan Lab), Stable Diffusion XL through InstantID (Team Brihadaranyaka), and FLUX.1-dev (Team Go-To-Germany).

Despite this diversity, the pipelines converged on a small set of shared components. Stable Diffusion 1.5 was the most frequently used backbone, and the combination of ControlNet for landmark and geometry conditioning with an IP-Adapter (most often IP-Adapter-FaceID) for identity injection appeared in nearly every diffusion pipeline. Frame-selection heuristics based on sharpness, frontality, and proximity to the target landmarks were also widely used to obtain a clean identity reference before generation.

Adversarial post-processing also emerged as a very common implementation among participants. Nearly all teams applied a projected gradient descent (PGD) attack against one or more surrogate detectors, frequently combined with JPEG compression, Gaussian blur, and additive noise to suppress the residual artifacts that frequency-domain detectors exploit. Despite that, many teams did not achieve very high evasion scores, with submissions ranging from 0.17 to 0.67 evasion against the detectors implemented by other participants. At the same time, some teams achieved significantly higher performance in evasion against the open-source detectors used by the organizers, with team Go-To-Germany achieving a score of 0.9019. About half of the teams focused on maintaining a small landmark distance versus the organizer-imposed constraints, while the majority of teams averaged a good identity similarity.

The two strongest submissions, Team Brihadaranyaka (overall score 0.4364) and Team Go-To-Germany (0.4170), succeeded because they balanced generation quality against evasion rather than optimizing one at the expense of the other. Brihadaranyaka’s unified design reused a single proxy detector both as a PGD target during generation and as an ensemble member during detection, ensuring that the adversarial perturbations were aligned with the detection objective; paired with the high-fidelity Stable Diffusion XL and InstantID backbone, this produced images that were realistic and, at the same time, difficult to flag. Go-To-Germany combined a strong FLUX.1-dev generator with an unusually sophisticated attack, a PGD perturbation jointly optimized against twelve white-box detectors and hardened with Input Diversity and Expectation over Transformation to transfer to unseen models. In both cases a capable generative backbone provided the quality while carefully engineered adversarial post-processing provided the evasion, and neither component alone would have been sufficient.

Finally, across submissions, image quality and evasion were positively correlated: higher-quality generations tended to evade the detectors more effectively. This suggests that the residual synthesis artifacts that degrade perceptual quality are largely the same cues the detectors rely on, so gains in realism and gains in evasion acted as complementary rather than competing objectives.

Overall, the generation sub-task proved demanding for many teams. This difficulty likely stems from the fact that participants were required to satisfy three competing objectives simultaneously: landmark-imposed generation, identity preservation, and evasion of deepfake detection. Balancing these constraints at once left little room for teams to specialize, and the results reflect that tension. We believe the task could be made more accessible in future editions, for instance by decoupling these objectives or introducing them in stages. At the same time, the strong participation demonstrates that the challenge resonates with the community and addresses a problem of genuine interest.

6.2.2. Detection

The detection sub-task drew an even wider range of approaches than generation, spanning semantic, spatial, frequency, and language-model signals. Semantic-feature detectors built on large pretrained encoders were the most common starting point: several teams fine-tuned CLIP-ViT-L/14, often adapting only the Layer Normalization parameters to preserve the encoder’s general representations, while others paired it with self-supervised backbones such as DINOv2 or DINOv3. A second family targeted low-level artifacts directly, including FFT frequency-ratio and spatial-frequency dual-stream detectors, neighboring-pixel-relation residuals, and high-frequency residual modeling with a Swin backbone. Beyond these, teams explored less conventional formulations: CLIP feature-space k-NN anomaly scoring and zero-shot text-prompt classification, Self-Blended Images for online pseudo-forgery generation, Cross-EfficientViT cross-attention, and the routing of uncertain samples to vision-language judges such as Qwen2.5-VL and Claude Sonnet 4.6.

Despite the variety of submission strategies and models, almost all of them shared a clear common structure. Almost every team relied on an ensemble rather than a single classifier, fusing complementary backbones through weighted averaging, max-probability selection, or a logistic-regression stacking meta-learner. CLIP and other ViT-family encoders were the dominant backbones, frequency-domain features recurred as a complementary signal to spatial cues, and FaceForensics++ was the most widely used training source, often supplemented with CelebDF-v2, DF40, or diffusion-generated faces. Test-time augmentation through flips, crops, and multi-scale views was also near-universal, and the use of a vision-language model as an auxiliary judge for ambiguous cases emerged as a notable recurring idea, although the teams that tried it reported little measurable gain.

The task nonetheless proved difficult. Only a small number of teams exceeded a weighted accuracy of 0.6, and many systems clustered well below this level. Given that the metric is a balanced weighted accuracy, scores in this region indicate performance close to random on one or both classes, underscoring how challenging robust generalization to the unseen generators in the test set remained. The two strongest submissions were Team Go-To-Germany (weighted accuracy 0.6987) and Team k20221289 (0.6936), both of which combined complementary encoders selected on the basis of a self-curated benchmark: Go-To-Germany paired a full-image AI-generation specialist with a face-manipulation specialist, while k20221289 fused a language-aligned CLIP detector aimed at GAN and FaceSwap artifacts with a self-supervised DINOv2 detector emphasized toward diffusion outputs.

A recurring failure mode across the leaderboard was threshold calibration. Because the decision threshold directly governs the trade-off between real-image and fake-image accuracy, teams that fixed it at 0.5 without validation-set calibration, or that calibrated it against a flawed criterion, tended to collapse toward one class: some achieved high accuracy on real images while missing most deepfakes, and others showed the opposite imbalance. A well-chosen threshold was often as consequential as the detector architecture itself, and that strong per-class accuracy did not guarantee a strong balanced score.

Finally, when the deepfake-detection accuracy was weighted by image quality, every team’s score dropped relative to the unweighted variant. This consistent decline indicates that higher-quality generations were systematically harder to detect: the quality weighting places more emphasis on the most realistic fakes, precisely the images on which the detectors were least reliable. The result mirrors the correlation observed in the generation sub-task, where higher-quality images also evaded detectors more effectively, and it confirms that perceptual realism and detectability are closely linked from both sides of the challenge.

7. Conclusion

This paper presents the ImageCLEF 2026 Deepfake Task overview. The task proposed two challenges: Deepfake Detection and Deepfake Generation, across two different modalities: images and audio. The paper presents the tasks, datasets, evaluation methods in detail and introduces the methods proposed by

the task participants. Finally, the paper introduces a discussion around the results in the benchmarking tasks and the challenges encountered in the field of deepfake generation and detection.

A common theme throughout this paper is the lack of generalization in deepfake models: the diversity in generation quality, methods and the unknown distribution of data are challenges that mirror real-life scenarios. Looking across both the image and audio modalities, there is a clear asymmetry between the two challenges: detection proved more tractable than generation, since detectors could draw on representation diversity through heterogeneous ensembles, while generators had to satisfy quality, identity, intelligibility, and evasion constraints at once.

For the audio modality, the first edition of the ImageCLEF 2026 Deepfake Task showed that few-shot voice cloning can already produce intelligible and moderately speaker-consistent synthetic speech, but that robust evasion remains difficult. Most generation systems relied on zero-shot or few-shot TTS models, especially XTTS v2 and F5-TTS, often combined with reference selection, model ensembling, speaker-similarity optimization, and signal-level post-processing. The best systems balanced transcript preservation, speaker similarity, naturalness, and evasion, while submissions that optimized only audio quality were limited by the multiplicative scoring scheme.

The main limitation on the generation side was detector evasion, especially against organizer-side systems. Although several teams achieved low WER and CER values and high WavLM speaker-similarity scores, post-processing methods such as filtering, codec simulation, noise injection, loudness normalization, or time stretching did not consistently transfer across detector families. The difference between WavLM and ECAPA-TDNN speaker-similarity scores also shows that identity preservation remains dependent on the evaluation representation, justifying the use of complementary speaker encoders.

For audio detection, the strongest results were obtained by heterogeneous ensembles combining self-supervised speech models and speaker-recognition features. Systems based on WavLM, Wav2Vec2/XLS-R, ECAPA-TDNN, x-vector embeddings, AASIST, and LFCC features generally performed better when combined through score fusion, chunk-based inference, self-augmentation, or calibration. The best detector achieved a final score of 0.9522 and perfect accuracy on participant-generated deepfakes, showing that robust audio detection is feasible when multiple artifact families are modeled. At the same time, several detectors showed clear class-specific biases, either misclassifying real recordings as fake or failing to detect participant-generated deepfakes. These failures were mainly linked to a distribution mismatch between external training data and the hidden evaluation set, as well as insufficient threshold calibration. Overall, the audio results indicate that detection was more successful than generation in this edition: detectors benefited from representation diversity, while generators had to satisfy quality, identity, intelligibility, and evasion constraints simultaneously.

On the image generation side, participants converged on diffusion-based synthesis with landmark and identity conditioning, coupled with increasingly sophisticated adversarial post-processing. Stable Diffusion 1.5 was the most common backbone, typically combining ControlNet for landmark and geometry conditioning with an IP-Adapter for identity injection, InsightFace embeddings for identity extraction, and MediaPipe FaceMesh landmarks as the structural signal, though a minority of teams instead framed the problem as warping or reenactment. The requirement to satisfy landmark fidelity, identity preservation, and detector evasion simultaneously made strong all-round performance rare, and the two best submissions succeeded precisely because they balanced generation quality against evasion rather than optimizing one at the expense of the other, pairing a capable generative backbone with carefully engineered adversarial post-processing. As in the audio modality, evasion was the hardest of the three objectives: nearly all teams applied projected gradient descent attacks against surrogate detectors, but these perturbations did not transfer reliably to the organizer-side systems, and the schemes that helped most were those that operated on the image actually seen by the detectors, such as differentiable JPEG simulation.

For image detection, ensemble systems built on large pretrained encoders dominated, most often fusing CLIP and other ViT-family backbones with complementary self-supervised or frequency-domain signals through weighted averaging or stacking. Only a few teams surpassed a weighted accuracy of 0.6, with many systems clustering below that level, indicating performance close to random on one or both

classes and underscoring how difficult robust generalization to the unseen test generators remained. Threshold calibration emerged as a recurring failure mode: several teams collapsed toward one class, achieving high accuracy on real images while missing most deepfakes or the reverse, and a poorly chosen threshold was often as consequential as the detector architecture itself. Across both image sub-tasks, image quality and detectability proved tightly coupled, with higher-quality generations both evading detectors more effectively and being harder to classify correctly, a pattern confirmed by the consistent drop in quality-weighted detection accuracy relative to the unweighted variant.

These findings mark the clearest direction for future work, both in building detectors that stay reliable against high-fidelity fakes and in refining the task design so that its combined objectives become more accessible to participants in future editions. Furthermore, future editions of the task plan to improve and enlarge the datasets, to add stricter submission rules, and to add a truly multimodal task, combining video and audio.

Acknowledgements

The work of Dan-Cristian Stanciu was funded by the Ministry of Research, Innovation and Digitization, UE-FISCDI, Solutions Axis, grant SOL4/2024, project number PN-IV-P6-6.3-SOL-2024-2-0320. The work of Mihai Gebriel Constantin was supported by UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0060, DeteRel, 1Sol(T12)/2024.

Declaration on Generative AI

During the preparation of this work, the authors used generative models (Claude Anthropic) in order to: enhance the grammar, spelling check and paraphrasing. All scientific content and references were provided by the authors. After using these tools, the authors reviewed and edited the content as needed and take 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] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, et al., Xtts: a massively multilingual zero-shot text-to-speech model, *arXiv preprint arXiv:2406.04904* (2024).
- [3] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi, et al., Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training, *arXiv preprint arXiv:2505.17589* (2025).
- [4] 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).
- [5] H. Kim, S. Kim, J. Yeom, S. Yoon, Unitspeech: Speaker-adaptive speech synthesis with untranscribed data, *arXiv preprint arXiv:2306.16083* (2023).

- [6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.
- [7] digiplay, Photon_v1, https://huggingface.co/digiplay/Photon_v1, 2023. Hugging Face model repository. Accessed: 2026-07-03.
- [8] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving latent diffusion models for high-resolution image synthesis, in: *International Conference on Learning Representations (ICLR)*, 2024.
- [9] SG161222, RealVisXL_V4.0, https://huggingface.co/SG161222/RealVisXL_V4.0, 2024. Hugging Face model repository. Accessed: 2026-07-03.
- [10] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: *Interspeech*, 2020.
- [11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, *IEEE Journal of Selected Topics in Signal Processing* 16 (2022) 1505–1518.
- [12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, in: *Proceedings of the 40th International Conference on Machine Learning*, 2023.
- [13] G. Mittag, B. Naderi, A. Chehadi, S. Möller, Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets, in: *Proc. Interspeech 2021*, 2021, pp. 2127–2131.
- [14] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, N. Evans, Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, *arXiv preprint arXiv:2202.12233* (2022).
- [15] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, B. Wu, Deepfakebench: A comprehensive benchmark of deepfake detection, 2023. URL: <https://arxiv.org/abs/2307.01426>. *arXiv*: 2307.01426.
- [16] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, FaceForensics++: Learning to Detect Manipulated Facial Images , in: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 1–11. URL: <https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00009>. doi:10.1109/ICCV.2019.00009.
- [17] F. Chollet, Xception: Deep learning with depthwise separable convolutions, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) 1800–1807. URL: <https://api.semanticscholar.org/CorpusID:2375110>.
- [18] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 6105–6114. URL: <https://proceedings.mlr.press/v97/tan19a.html>.
- [19] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (2018) 1–7. URL: <https://api.semanticscholar.org/CorpusID:52157475>.
- [20] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, Springer-Verlag, Berlin, Heidelberg, 2020, p. 86–103. URL: https://doi.org/10.1007/978-3-030-58610-2_6. doi:10.1007/978-3-030-58610-2_6.
- [21] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, N. Yu, Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain , in: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 772–781. URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00083>. doi:10.1109/CVPR46437.2021.00083.
- [22] Y. Wu, X. Song, J. Chen, Y.-G. Jiang, Generalizing face forgery detection via uncertainty learning, in: *Proceedings of the 31st ACM International Conference on Multimedia, MM '23*, Association for Computing Machinery, New York, NY, USA, 2023, p. 1759–1767. URL: <https://doi.org/10.1145/3581783.3612102>. doi:10.1145/3581783.3612102.

- [23] Z. Yan, Y. Zhang, Y. Fan, B. Wu, Ucf: Uncovering common features for generalizable deepfake detection, 2023. URL: <https://arxiv.org/abs/2304.13949>. arXiv:2304.13949.
- [24] Y.-S. Ni, D. Meng, C. Yu, C. Quan, D. Ren, Y. Zhao, Core: Consistent representation learning for face forgery detection, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022) 12–21. URL: <https://api.semanticscholar.org/CorpusID:249395469>.
- [25] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, X. Yang, End-to-End Reconstruction-Classification Learning for Face Forgery Detection, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2022, pp. 4103–4112. URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00408>. doi:10.1109/CVPR52688.2022.00408.
- [26] D.-C. Stanciu, B. Ionescu, Improving generalization in deepfake detection via augmentation with recurrent adversarial attacks, in: Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, MAD '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 46–54. URL: <https://doi.org/10.1145/3643491.3660291>. doi:10.1145/3643491.3660291.
- [27] D.-C. Stanciu, B. Ionescu, Autoencoder-based data augmentation for deepfake detection, in: Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation, MAD '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 19–27. URL: <https://doi.org/10.1145/3592572.3592840>. doi:10.1145/3592572.3592840.
- [28] Y. Li, D. Zhu, X. Cui, S. Lyu, Celeb-df++: A large-scale challenging video deepfake benchmark for generalizable forensics, ArXiv abs/2507.18015 (2025). URL: <https://api.semanticscholar.org/CorpusID:280017220>.
- [29] S. Kim, J. Kim, J. Woo, Go-To-Germany at ImageCLEF 2026 Audio Deepfake Detection and Generation: A Multi-Backbone SSL Detection System with Cross-Track Insights, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [30] Anubhav, Brihadaranyaka at imageclef-deepfake 2026: A unified system for generation and detection with anchor-aligned ensemble calibration, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [31] S. Kim, Y. Lee, J. Kim, Chaboom at ImageCLEF 2026 Deepfake Task: Multimodal Deepfake Detection and Generation, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [32] S. Park, W. Kim, Y. Lee, Hybrid Few-Shot Generation and Multi-Model Ensemble Detection of Deepfake Faces and Voices, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [33] S. A. Tarimo, E. E. Ngasa, G. L. Matrin, Self-augmented audio-visual deepfake detection with zero-shot voice cloning and face reenactment, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [34] J. Kang, Multimodal Deepfake Detection Using Heterogeneous Vision Transformer Ensemble with Face Quality Estimation, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [35] D. Emakporuena, R. N. Chowdhury, S. S. Alam, E. P. O. Oluwafemi, M. Hoque, O. G. Akingbola, M. M. Rahman, Cross-architecture detection and constraint-guided generation for audio-visual deepfake evaluation: Lessons from imageclef 2026, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [36] Q. Hu, Z. Han, Z. Su, J. Peng, W. Huang, W. Huang, Face-Aligned Image Ensembles and LFCC Audio Transformers for ImageCLEF 2026 Deepfake Detection, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [37] C. Cheng, Z. Xiaobing, Multi-feature audio fusion and high-frequency image modeling, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
- [38] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, X. Chen, F5-tts: A fairytaler 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), 2025, pp. 6255–6271.

- [39] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, F. Wei, Speech5: Unified-modal encoder-decoder pre-training for spoken language processing (2021). [arXiv:2110.07205](https://arxiv.org/abs/2110.07205).
- [40] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, *Advances in neural information processing systems* 33 (2020) 17022–17033.
- [41] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, et al., Speechbrain: A general-purpose speech toolkit, *arXiv preprint arXiv:2106.04624* (2021).
- [42] J. Cui, Z. Yang, N. Li, J. Tian, X. Ma, Y. Zhang, G. Chen, R. Yang, Y. Cheng, Y. Zhou, G. Yu, X. Gu, J. Tang, GLM-TTS Technical Report, in: *arXiv 2512.14291*, 2025.
- [43] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, Z.-H. Ling, ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech, in: *arXiv 1911.01601*, 2019.
- [44] MelodyMachine, Deepfake-audio-detection-V2, Hugging Face model card, 2024. URL: <https://huggingface.co/MelodyMachine/Deepfake-audio-detection-V2>, audio classification model for deepfake audio detection.
- [45] Hemgg, Deepfake-audio-detection, Hugging Face model card, 2024. URL: <https://huggingface.co/Hemgg/Deepfake-audio-detection>, the Heem2 URL redirects to Hemgg/Deepfake-audio-detection.
- [46] Gustking, wav2vec2-large-xlsr-deepfake-audio-classification, Hugging Face model card, 2024. URL: <https://huggingface.co/Gustking/wav2vec2-large-xlsr-deepfake-audio-classification>, wav2Vec2 model fine-tuned for deepfake audio classification.
- [47] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino, et al., Xls-r: Self-supervised cross-lingual speech representation learning at scale, *arXiv preprint arXiv:2111.09296* (2021).
- [48] G. Stafford, Deepfake audio detection dataset (v4), Hugging Face dataset, 2025. Dataset: [garystafford/deepfake-audio-detection](https://huggingface.co/datasets/garystafford/deepfake-audio-detection); License: CC-BY-4.0.
- [49] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, N. Evans, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371. doi:10.1109/ICASSP43922.2022.9747766.
- [50] J. Kim, S. Kim, J. Woo, Adversarial Deepfake Generation and an Investigation of Purification-Based Adversarial Detection, in: *CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [51] R. Biswas, W. Zaghouani, MarsadLab at ImageCLEF 2026 Deepfake Detection and Generation: Warp-Based Image Generation and Ensemble Deepfake Detection, in: *CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [52] R. Solanki, A Multi-Architecture Ensemble with Test-Time Augmentation and Adaptive Batch Normalization for Image Deepfake Detection, in: *CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [53] R. Jayasree, P. Mirunalini, M. K. A., A dual-stream spatial-frequency feature learning framework for generalized deepfake detection, in: *CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [54] D. A. Coccomini, N. Messina, C. Gennaro, F. Falchi, Combining EfficientNet and Vision Transformers for Video Deepfake Detection, Springer International Publishing, 2022, p. 219–229. URL: http://dx.doi.org/10.1007/978-3-031-06433-3_19. doi:10.1007/978-3-031-06433-3_19.
- [55] X. Wang, Y. Li, H. Zhang, Y. Shan, Towards real-world blind face restoration with generative facial prior, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9168–9178.

- [56] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann, Real-time facial surface geometry from monocular video on mobile gpus, arXiv preprint arXiv:1907.06724 (2019).
- [57] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836–3847.
- [58] 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).
- [59] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502 (2020).
- [60] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).
- [61] 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).
- [62] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.
- [63] SG_161222, Realistic vision V6.0 B1, https://huggingface.co/SG161222/Realistic_Vision_V6.0_B1_noVAE, 2024. Stable Diffusion 1.5 fine-tuned checkpoint. Available on Civitai: <https://civitai.com/models/4201/realistic-vision-v60-b1>. Accessed: 2026-07-05.
- [64] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
- [65] Black Forest Labs, FLUX.1-dev: An open-source diffusion transformer for image generation, <https://huggingface.co/black-forest-labs/FLUX.1-dev>, 2024. Diffusion Transformer model used for face generation. Accessed: 2026-07-05.
- [66] Shakker Labs, FLUX.1-dev ControlNet union pro 2.0, <https://huggingface.co/Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro-2.0>, 2024. Pretrained on OpenPose input convention. Accessed: 2026-07-05.
- [67] Z. Guo, Y. Wu, Z. Chen, L. Chen, P. Zhang, Q. He, Pulid: Pure and lightning id customization via contrastive alignment, Advances in neural information processing systems 37 (2024) 36777–36804.
- [68] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples, in: International conference on machine learning, PMLR, 2018, pp. 284–293.
- [69] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, D. Zhang, Liveportrait: Efficient portrait animation with stitching and retargeting control, arXiv preprint arXiv:2407.03168 (2024).
- [70] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Computer Graphics and Applications 21 (2001) 34–41. doi:10.1109/38.946629.
- [71] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: P. S. Heckbert (Ed.), Graphics Gems IV, Academic Press, 1994, pp. 474–485.
- [72] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (1981) 381–395. doi:10.1145/358669.358692.
- [73] A. Yermakov, J. Cech, J. Matas, Unlocking the hidden potential of clip in generalizable deepfake detection, ArXiv abs/2503.19683 (2025). URL: <https://api.semanticscholar.org/CorpusID:277313445>.
- [74] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE signal processing letters 23 (2016) 1499–1503.
- [75] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
- [76] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11975–11986.
- [77] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, et al., Df40: Toward next-generation deepfake detection, Advances in Neural Information Processing Systems 37 (2024)

29387–29434.

- [78] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
- [79] K. Shiohara, T. Yamasaki, Detecting deepfakes with self-blended images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18720–18729.
- [80] A. Yermakov, et al., Deepfake detection that generalizes across benchmarks, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2026.
- [81] BitMind, BitMind dataset, <https://huggingface.co/bitmind>, 2024. Dataset for AI-generated image detection. Accessed: 2026-07-05.
- [82] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.
- [83] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. Canton Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv:2006.07397 (2020).