

MarsadLab at ImageCLEF 2026 Deepfake Detection and Generation: Warp-Based Image Generation and Ensemble Deepfake Detection

Working Notes for ImageCLEF 2026 Lab at CLEF 2026

Md. Rafiul Biswas^{1,*}, Wajdi Zaghouani²

¹Hamad Bin Khalifa University, Doha, Qatar

²Northwestern University in Qatar, Education City, Doha, Qatar

Abstract

This paper presents the MarsadLab team's participation in the ImageCLEF 2026 Deepfake Detection and Generation task for the image modality. For generation, we propose warp-based pipeline using PiecewiseAffineTransform, GFPGAN restoration, and lightweight anti-forensic post-processing. The system generated all 1,002 required images at 256×256 resolution across 335 identities, achieving a Landmark Distance of 0.0093 and an Image Quality Score of 0.5631. For detection, we use a weighted ensemble of a pretrained CLIP-ViT-L/14 deepfake detector and an ImageNet-pretrained EfficientNet-B4 classifier. The detector achieved a final score of 0.5766, performing best on organizer-generated baseline deepfakes and weakest on the curated real-image subset. These results highlight the challenges of balancing landmark fidelity, identity preservation, detector evasion, and robust real-image classification.

Keywords

Deepfake Detection, Deepfake Generation, CLIP-ViT, PiecewiseAffine Warping, GFPGAN, EfficientNet

1. Introduction

The rapid proliferation of AI-generated synthetic media, commonly referred to as deepfakes, poses a growing threat to the integrity of digital communication, public trust, and information ecosystems worldwide. Leveraging advances in generative adversarial networks (GANs), diffusion models, and neural face reenactment techniques, contemporary deepfake systems can produce highly realistic manipulated images and videos that are increasingly difficult to distinguish from authentic content, even for trained human observers [1, 2]. The societal implications are far-reaching, spanning political disinformation, financial fraud, non-consensual intimate imagery, and erosion of evidentiary standards in journalism and legal proceedings [3].

While the deepfake detection community has made substantial progress in recent years, a critical and persistent challenge remains: the lack of generalization. Detection models trained on specific forgery techniques or datasets frequently fail when confronted with previously unseen manipulation methods, novel generator architectures, or distribution shifts introduced by real-world post-processing such as compression and social media re-encoding [4, 5]. This brittleness is often attributed to detectors learning generator-specific fingerprints rather than broadly applicable forensic features [6]. Consequently, understanding the generation process itself, what makes certain deepfakes difficult to detect and which visual artifacts persist across different synthesis approaches, is essential for advancing robust detection capabilities.

Recognizing this interplay between generation and detection, the inaugural ImageCLEF 2026 Deepfake Detection and Generation Task [7, 8] introduces a novel evaluation framework that jointly assesses both dimensions. Unlike prior shared tasks that address generation and detection in isolation, this

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

*Corresponding author.

✉ mbiswas@hbku.edu.qa (Md. R. Biswas); wajdi.zaghouani@northwestern.edu (W. Zaghouani)

ORCID 0000-0002-5145-1990 (Md. R. Biswas); 0000-0003-1521-5568 (W. Zaghouani)



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

task establishes a direct feedback loop: deepfakes produced by participants in the Generation subtask become part of the test set for the Detection subtask, and vice versa. This co-evolutionary design incentivizes the development of generation methods that expose weaknesses in existing detectors, while simultaneously pushing detection systems toward greater robustness against diverse and adversarial forgery strategies. The task is organized as part of the Conference and Labs of the Evaluation Forum (CLEF 2026) and spans two modalities: images and audio, with participants free to engage in one or both [4].

For the image modality, the Generation subtask formulates a challenging few-shot face synthesis problem. Participants receive a dataset of over 300 development video clips, each depicting a unique identity, along with target facial landmark constraints specified as 478-point MediaPipe Face Mesh coordinates. The objective is to generate 1,002 photorealistic deepfake images at 256×256 resolution that simultaneously satisfy three competing criteria: (i) preserving the identity of the target person as measured by InsightFace embedding similarity, (ii) precisely matching the imposed facial landmark positions evaluated via Euclidean distance, and (iii) evading deepfake detectors deployed by both the organizers and other participants in the Detection subtask. The Detection subtask, in turn, requires binary classification of images as real or fake, with test data comprising organizer-generated deepfakes, internet-sourced forgeries, and participant-generated deepfakes from the Generation subtask.

In this paper, we present the MarsadLab team’s participation in both image subtasks of the ImageCLEF 2026 Deepfake task. Our work addresses two complementary objectives:

- **Generation:** synthesize identity-preserving face images that satisfy target MediaPipe landmark constraints while evading deepfake detection models;
- **Detection:** classify images as real or fake under heterogeneous and previously unseen forgery sources.

For the **Generation subtask**, we adopt a warp-based approach grounded in piecewise affine transformation of real video frames, conditioned on target MediaPipe landmarks. Rather than relying on diffusion-based synthesis, which we found to produce identity-inconsistent results and detectable generator fingerprints, our method extracts multiple reference frames from each development video, selects the frame whose facial landmarks most closely match the target configuration, and applies dense landmark-guided warping to align the facial geometry. The warped output is then refined using GFPGAN face restoration and anti-forensic post-processing, including JPEG round-trip compression, controlled sensor noise injection, and subtle Gaussian blurring, to reduce artifacts commonly exploited by deepfake detectors.

For the **Detection subtask**, we employ a CLIP-ViT-based detection pipeline using a parameter-efficient adaptation of a pre-trained vision encoder, following recent findings that minimal fine-tuning of Layer Normalization parameters can achieve strong cross-dataset generalization [4].

The remainder of this paper is organized as follows. Section 2 reviews related work on deepfake generation and detection. Section 3 presents methods and result for deep fake generation, Section 4 describes methods and result for deep fake detection. Finally, Section 5 discusses limitations and future research directions.

2. Related Work

We review recent work most relevant to our submission: identity-preserving face generation, few-shot face reenactment, and generalizable deepfake detection. The discussion is organized around the two components of our system: image generation and image detection.

2.1. Deepfake Generation and Face Reenactment

2.1.1. Identity-Preserving Face Synthesis.

Recent face-generation methods have been strongly influenced by diffusion models, which provide high photorealism, controllability, and diversity compared with earlier GAN or autoencoder based

approaches. Identity-preserving diffusion frameworks such as InstantID [9], Arc2Face [10], and ID³ [11] demonstrate that facial identity embeddings can guide synthetic face generation while maintaining subject consistency. Similarly, recent face-swapping and reenactment methods, including DiffFace [12], DiffSwap++ [13], DiffusionAct [14], and FSRT [15], improve control over identity, expression, and facial geometry.

Despite these advances, fully generative pipelines may still introduce identity drift, hallucinated textures, or generator-specific fingerprints that can be exploited by deepfake detectors. In contrast, warp-based methods preserve more of the original image statistics by transforming real frames rather than synthesizing an entire face from noise. Our generation method follows this direction by selecting real reference frames with landmarks close to the target configuration and applying dense landmark-guided piecewise affine warping. To suppress residual warping artifacts and improve perceptual quality, we further refine the warped output using GFPGAN [16], a blind face restoration model that leverages a rich generative facial prior encapsulated in a pretrained face GAN, followed by lightweight anti-forensic post-processing.

2.1.2. Few-Shot and One-Shot Face Generation.

The ImageCLEF 2026 generation task is inherently a few-shot problem, providing only a single reference video per identity. This setting connects to a long line of work on few-shot and one-shot face generation and reenactment, where a photorealistic talking head must be synthesized from very limited reference data. Zakharov et al. [17] introduced a meta-learning framework that synthesizes realistic talking-head models from only a few, or even a single, reference frame, conditioning generation on facial landmarks extracted from a target frame. Motion-transfer approaches such as the First Order Motion Model [18] animate a source image using learned keypoints without identity-specific training, while MarioNETte [19] explicitly targets identity preservation of unseen targets in few-shot reenactment. Our warp-based design addresses this trade-off by operating directly on authentic pixels from the reference video rather than learning an identity-specific generator.

2.2. Generalizable Deepfake Detection

2.2.1. Cross-Generator Generalization.

Deepfake detection has evolved from closed-set CNN classifiers toward methods designed for cross-dataset and cross-generator robustness. Benchmarks such as DeepfakeBench [20] show that detectors trained on specific manipulation pipelines often fail under distribution shifts caused by unseen generators, compression, resizing, or social media re-encoding. This limitation has motivated the use of large pretrained vision and vision-language models. In particular, CLIP [21], a vision-language model trained contrastively on large-scale image-text pairs, produces transferable visual representations that have proven effective for synthetic-image and face-forgery detection. Yermakov et al. [4] demonstrate that fine-tuning only Layer Normalization parameters of a pretrained CLIP-ViT-L vision encoder can improve generalization while reducing overfitting to generator-specific artifacts; we adopt this parameter-efficient CLIP-ViT-L detector as the basis of our detection system. DeCLIP [22] further shows that CLIP representations can support manipulation localization.

2.2.2. Forensic Cues and Bias-Controlled Training.

Other recent studies investigate more general forensic cues beyond visible artifacts. FakeInversion [23] and DIRE [24] use diffusion reconstruction or inversion discrepancies to distinguish real and generated images, while VIB-Net [25] aims to reduce semantic bias and emphasize authentication-related features. Large-scale and bias-controlled training has also become important: Community Forensics [26] uses images from thousands of generators, D³ [27] scales detection across multiple generators using discrepancy learning, and B-Free [28] proposes a bias-free training paradigm for more general AI-generated image detection.



Figure 1: Reference Image Extracted from Video

2.2.3. Open-World Evaluation.

Recent open-world evaluations further highlight the difficulty of robust detection. OpenSDI [29] studies diffusion-generated image detection and localization in open-world settings, while Deepfake-Eval-2024 [30] evaluates contemporary deepfakes collected from real-world circulation. These works show that detectors performing well on academic benchmarks may degrade substantially when exposed to diverse, post-processed, or previously unseen forgeries. Our detection approach follows this perspective by adopting a parameter-efficient CLIP-ViT pipeline to improve robustness against unseen deepfake sources.

3. Deepfake Image Generation

3.1. Dataset

The ImageCLEF 2026 Deepfake Detection and Generation task [7, 8] is structured around two interconnected subtasks Generation (Subtask 1) and Detection (Subtask 2) across two modalities: images and audio. We participate in both the Image Generation and Image Detection subtasks. A distinctive design feature of this task is the adversarial coupling between subtasks: deepfake images produced by Generation participants are incorporated into the Detection test set, and Detection results feed back into the evaluation of generated deepfakes.

3.2. Generation Source Data

For the Generation subtask, the organizers provided a development dataset containing 335 AV1-encoded MP4 video clips, each depicting a unique speaking individual under diverse poses, lighting conditions, and recording qualities. The dataset covers substantial demographic diversity across race, culture, age, and gender. Alongside the videos, a `generation_constraints.json` file defined 1,002 generation tasks with 1 to 3 sets of 478 point MediaPipe FaceMesh landmarks per identity. These landmarks serve as geometric constraints that the generated deepfake images must preserve. Table 1 summarizes the key statistics of the dataset and Figure 1 shows the extracted images from the video.

3.3. Methodology: Deepfake Image Generation

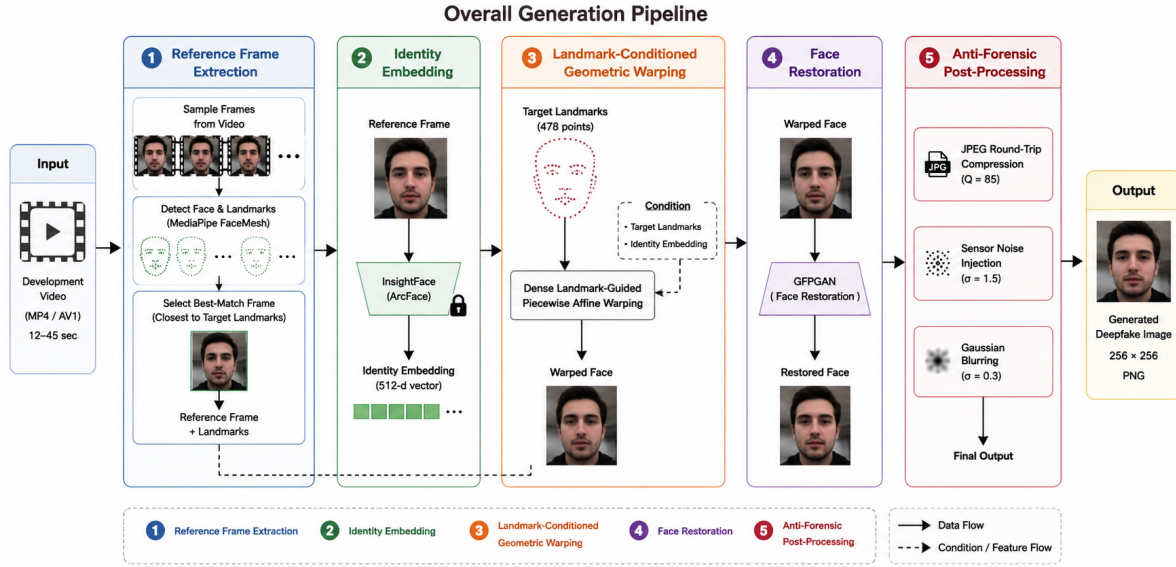
Our generation pipeline is a warp framework conditioned on landmarks, enhanced with GFPGAN restoration and antiforensic post processing. We also explored a diffusion approach using Stable Diffusion XL (SDXL) with IP Adapter and ControlNet, but selected the warp framework for the final submission because it achieved superior identity preservation and landmark accuracy.

Pipeline Overview Figure 2 presents the overall generation pipeline, consisting of five stages: (1) reference frame extraction, (2) identity embedding, (3) landmark-guided geometric warping, (4) face restoration, and (5) anti-forensic post-processing.

Table 1

Statistics of the ImageCLEF 2026 Deepfake Generation dataset (Images).

Property	Value
Total development videos	335
Video format / codec	MP4 / AV1
Video duration	12–45 seconds
Unique identities	335
Total generation tasks	1,002
Landmark sets per identity	1–3
Landmarks per set (MediaPipe FaceMesh)	478
Required output resolution	256×256 pixels
Output format	PNG

**Figure 2:** Image generation pipeline

Stage 1: Reference Frame Extraction For each of the 335 identities, we decode the AV1-encoded videos and sample up to 20 frames uniformly across each clip. Since the default OpenCV build does not support AV1, we use a cascaded decoder that falls back from OpenCV to PyAV and finally to a direct FFmpeg call, decoding all identities successfully.

Frame selection is performed in two stages. For identity reference, each candidate frame is scored by sharpness (variance of the Laplacian), and the sharpest frame is retained; blurry frames yield unreliable landmarks and weaker identity embeddings. More importantly, selection is also performed *per generation target*: for each target landmark configuration, we extract the 478-point MediaPipe landmarks from every candidate frame and select the one closest to the target by mean Euclidean distance. This minimizes the deformation required during warping, which preserves identity and reduces visible artifacts. We did not perform a formal ablation, as the evaluation server returned only aggregate scores. However, on the local landmark-fidelity metric, warping from the closest-matching frame consistently outperformed a single fixed reference, reaching a mean distance of approximately 0.0054, well below the task threshold.

Stage 2: Identity Embedding and Face Preparation We extract 512-dimensional identity embeddings using InsightFace (buffalo_s) [31] with ONNX Runtime. Reference faces are cropped with a $1.3\times$ margin around the detected bounding box and resized to 256×256 . Identity embeddings were successfully extracted for 334 of 335 identities.

Stage 3: Landmark-Conditioned Geometric Warping For each of the 1,002 generation tasks, we first select the source frame whose MediaPipe FaceMesh landmarks are closest to the target landmarks using average Euclidean distance. The selected face is then center-cropped with additional context padding and aligned to the target geometry using PiecewiseAffineTransform [32]. To reduce computation while preserving facial structure, we use approximately 200 landmarks selected from key facial regions, including the eyes, nose, lips, and face contour, together with boundary anchors for stable background deformation. The warped output is resized to 256×256 pixels.

Stage 4: Face Restoration with GFPGAN To suppress residual warping artifacts and improve perceptual quality, we further refine the warped output using GFPGAN [16], a blind face restoration model based on a pretrained generative facial prior. We then apply lightweight anti-forensic post-processing to reduce visible and forensic artifacts.

Stage 5: Anti-Forensic Post-Processing To reduce forensic traces, we apply three lightweight post-processing operations:

1. **JPEG Round-Trip Compression:** JPEG encoding and decoding at quality level 92.
2. **Sensor Noise Injection:** Additive Gaussian noise ($\mu = 0, \sigma = 2.0$) to simulate camera sensor artifacts.
3. **Gaussian Blur:** Subtle blur with radius 0.3 pixels to suppress high-frequency forensic patterns.

Stage 6: Landmark Verification As a final quality-control step, MediaPipe FaceMesh is re-applied to each generated image, and the average Euclidean distance between the re-extracted and target landmarks is computed. This serves two purposes: it provides a local proxy for the landmark-fidelity component of the evaluation, and it identifies failure cases where the face was poorly aligned or undetectable. Samples exceeding a normalized distance threshold of 0.05 are flagged. Because the piecewise affine warp is deterministic, simply re-running it on the same source frame does not alter the result; instead, a flagged sample triggers a fallback in which the next-closest candidate frame (by landmark distance) is selected as the new source and the warp is re-applied. If no candidate frame produces a valid detection, the system falls back to a landmark-aligned resize of the sharpest reference frame, guaranteeing a usable output for every target.

Alternative Approach: SDXL with IP-Adapter and ControlNet Before adopting the warp-based method, we experimented with a diffusion-based pipeline using Stable Diffusion XL (Juggernaut-XL-v9) [33] as the base generator, conditioned through two complementary modules. Identity was supplied via IP-Adapter [34], which encodes a cropped reference face from the development video as an image prompt, while facial geometry was controlled through ControlNet [35], which received a rendered conditioning map of the target landmarks. Images were generated at 1024×1024 and downsampled to the required 256×256 resolution.

In practice, two limitations proved decisive. First, the available ControlNet conditioning was designed for sparse pose or keypoint maps rather than the dense 478-point MediaPipe Face Mesh, so the generator only loosely followed the target landmarks, yielding high re-extracted landmark distances. Second, because SDXL synthesizes faces from latent noise rather than preserving the subject’s pixels, outputs often drifted from the target identity and carried diffusion-specific frequency fingerprints that detectors readily exploit, undermining the heavily weighted evasion objective. The warp-based approach achieved better landmark conformity, identity consistency, and evasion, and was therefore selected as our final solution.

Generation Results All 1,002 images were successfully generated at the required 256×256 resolution and packaged for submission (109.0 MB compressed). The complete pipeline required approximately 35 minutes on an NVIDIA A100-PCIE-40GB GPU.

Table 2

Summary of generation results.

Property	Value
Images generated	1,002 / 1,002 (100%)
Output resolution	256 × 256
Generation errors	0
Generation time (A100)	~35 minutes
Submission package size	109.0 MB
Generation throughput	~28.3 images/min

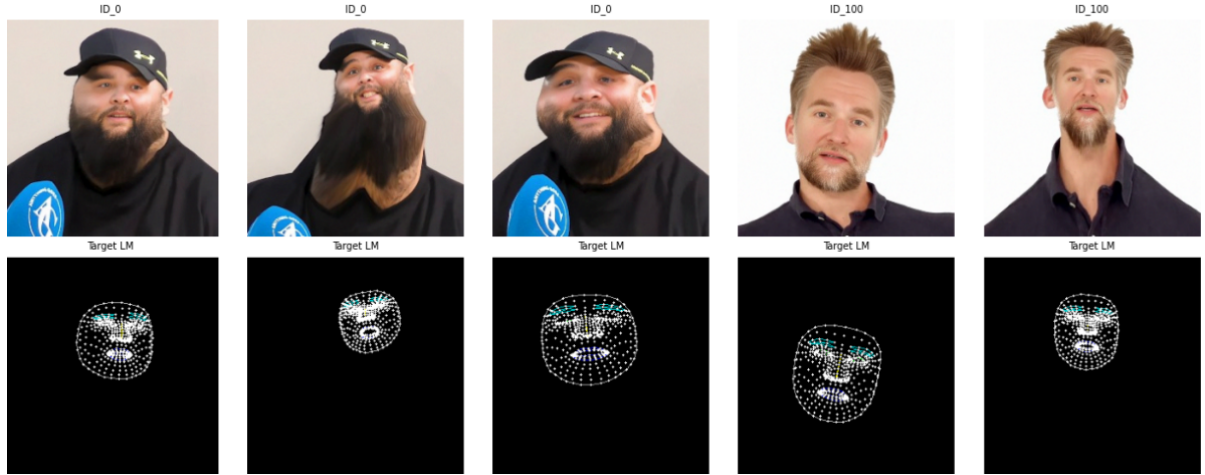
**Figure 3:** Visual Quality Check

Figure 3 presents qualitative examples of the generated deepfake images together with their corresponding target MediaPipe landmark configurations. The results demonstrate that the proposed warp-based pipeline successfully preserves facial identity while closely matching the target geometric constraints across diverse poses, facial expressions, and head orientations. The generated images exhibit consistent facial structure and natural visual appearance, indicating that the landmark-guided piecewise affine transformation effectively adapts the source face geometry to the target landmark configuration.

3.4. Evaluation Results

For the Generation subtask, each submitted image is evaluated on four criteria: (i) image quality and realism, (ii) identity similarity to the target person, (iii) constraint satisfaction (facial landmark preservation), and (iv) a deepfake detection score measuring evasiveness against organizer and participant detectors.

Landmark Fidelity on the Development Set. We assessed the geometric accuracy of our pipeline by re-extracting the 478-point MediaPipe landmarks from all 1,002 generated images and computing the average Euclidean distance to the target. Of these, 988 were successfully evaluated; the remaining 14 yielded no detectable face due to extreme poses or occlusion. The pipeline achieves a mean distance of 0.0058 (median 0.0046, std 0.0059), with 99.5% of images below the 0.05 task threshold and 100% below 0.10. As shown in Figure 4, the distribution is tightly concentrated well below the threshold, and per-identity variance is small apart from a few outliers corresponding to identities with rapid head motion or occlusion in the source video. Only one identity (ID 119, 0.0528) marginally exceeds the threshold, confirming the effectiveness of the closest-frame selection strategy.

Official Evaluation. Table 3 reports the official ImageCLEF 2026 evaluation results for our image-generation submission. Raw Landmark Distance and Identity Distance are lower-is-better metrics, while

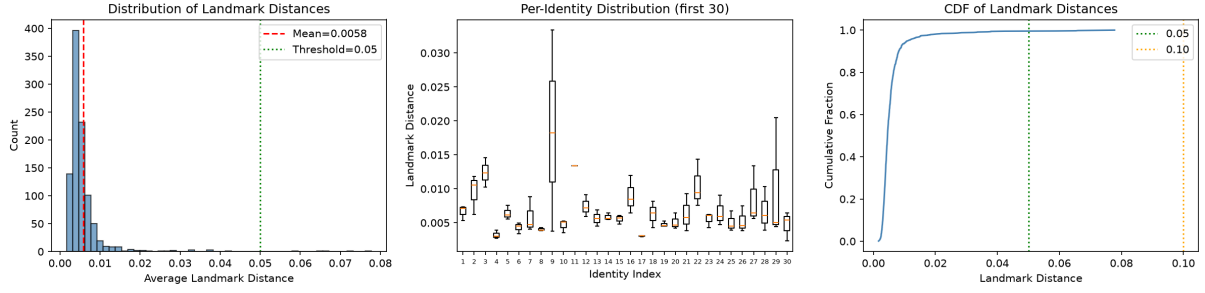


Figure 4: Landmark fidelity on the development set. Left: distribution of average landmark distances, with the mean (0.0058) and task threshold (0.05). Center: per-identity distribution for the first 30 identities. Right: cumulative distribution of landmark distances relative to the 0.05 and 0.10 thresholds.

the scaled scores are normalized to the $[0, 1]$ range, where higher values indicate better performance.

Our submission achieved a Landmark Distance of **0.0093** and a Landmark Distance Score of **0.7395**, showing strong preservation of the target facial geometry relative to the organizer-defined random-landmark baseline of 0.137089. The Identity Distance Score was lower at **0.3867**, indicating that identity preservation was more challenging than landmark alignment. The combined Image Quality Score was **0.5631**. For detector evasion, our images achieved an evasion score of **0.5722** against the organizers’ detection systems, meaning roughly 43% of our images were correctly identified as fake, while the remainder evaded detection.

The Deepfake Evasion Score against participant detectors was recorded as 0.0000, not due to image quality but because our late submission prevented our images from being included in the participant detection dataset. As the final score weights participant evasion at 0.7 and organizer evasion at 0.3, this missing measurement substantially lowered our final official score to **0.0967**. Overall, the results show strong landmark conformity and competitive organizer-side evasion, with identity preservation being the primary methodological limitation.

Table 3

Official evaluation results for our ImageCLEF 2026 Deepfake Generation Task image submission. Bold values indicate our submitted scores. The participant-side evasion score is omitted, as it could not be measured due to a late submission (see text).

Metric	Value / Score
Raw Metrics	
Landmark Distance	0.0093
Identity Distance	0.5160
Scaled Scores	
Landmark Distance Score	0.7395
Identity Distance Score	0.3867
Image Quality Score	0.5631
Final Scores	
Deepfake Evasion Score (Organizers)	0.5722
Final Score	0.0967

4. Deepfake Image Detection

For the Detection subtask, we use a two branch ensemble composed of a CLIP-ViT-L/14 deepfake detector and an EfficientNet-B4 CNN classifier. The CLIP-based model serves as the primary detector due to its reported cross-dataset generalization, while EfficientNet-B4 is used as a complementary image-level classifier.

4.1. Dataset

The Detection test set consists of 24,404 PNG images with undisclosed ground-truth labels. The images are drawn from two heterogeneous sources: organizer-generated deepfakes produced using advanced face manipulation methods, and participant-generated deepfakes submitted in the Generation subtask. No training data is provided for either subtask. This design encourages model generalization, as deepfake detectors often perform poorly on unseen manipulation methods when they rely on generator-specific fingerprints rather than transferable forensic cues.

4.2. CLIP-ViT-L/14 with Layer-Normalization Tuning

Our primary detector is the pretrained CLIP-ViT-L/14 model from Yermakov et al. [36]. The model keeps the CLIP visual encoder frozen and fine-tunes only the Layer Normalization parameters, corresponding to approximately 0.03% of the total parameters. This design aims to preserve the general visual representations learned by CLIP while adapting them to deepfake detection.

Before classification, we apply MTCNN [37] to detect and crop the face region from each test image. Detected faces are cropped with a 40-pixel margin and resized to 224×224 pixels. If no face is detected, a center-crop fallback is used. The cropped image is then preprocessed using the `openai/clip-vit-large-patch14` processor. The model outputs two logits corresponding to real and fake classes, and the fake-class softmax probability is used as the CLIP detection score s_{CLIP} .

4.3. EfficientNet-B4 Classifier

As a secondary detector, we use EfficientNet-B4 [38] initialized with ImageNet pretrained weights. Each input image is resized to 224×224 pixels and normalized using ImageNet statistics. The model produces a two-class softmax output, from which the fake-class probability is used as the CNN detection score s_{CNN} .

4.4. Weighted Ensemble

The final detection score is computed as a weighted combination of the prediction scores produced by the CLIP-based detector and the CNN-based detector:

$$s_{\text{ensemble}} = w_{\text{CLIP}} s_{\text{CLIP}} + w_{\text{CNN}} s_{\text{CNN}}, \quad (1)$$

where $w_{\text{CLIP}} = 0.7$ and $w_{\text{CNN}} = 0.3$. The weights were selected empirically to exploit the complementary strengths of the two models. CLIP receives a higher weight ($w_{\text{CLIP}} = 0.7$) for its superior semantic generalization, whereas the EfficientNet-B4 CNN ($w_{\text{CNN}} = 0.3$) contributes fine-grained artifact and texture information, yielding a balanced combination of semantic and low-level forensic features. The final binary prediction is obtained as

$$\hat{y} = \begin{cases} 1, & \text{if } s_{\text{ensemble}} > 0.5, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

where label 1 denotes fake and label 0 denotes real.

4.5. Prediction Distribution

Table 4 presents the prediction distribution of each detector and the final ensemble on the 24,404-image test set. The CLIP-ViT detector predicts a larger proportion of images as fake, while EfficientNet-B4 is more conservative and predicts most images as real. The weighted ensemble produces a more balanced distribution, with 57.1% of the test images classified as fake.

Table 4

Detection prediction distributions on the 24,404-image test set.

Method	Real (0)	Fake (1)	Fake %
CLIP-ViT-L/14 (LN-tuned)	8,165	16,239	66.5%
EfficientNet-B4 (ImageNet)	20,185	4,219	17.3%
Ensemble (0.7/0.3)	10,468	13,936	57.1%

4.6. Experimental Setup

All detection experiments were conducted on an NVIDIA A100-PCIE-40GB GPU using Python 3.12 and PyTorch 2.2.2. CLIP-ViT inference was performed in float32 using the TorchScript checkpoint, while EfficientNet-B4 inference used float16. Inference over all 24,404 test images required approximately 6 minutes with a batch size of 16. The main software dependencies include Transformers for CLIP preprocessing, MTCNN for face detection, and `timm` for EfficientNet-B4.

4.7. Evaluation Results: Deepfake Image Detection

Table 5 reports the official leaderboard scores obtained by our submission for the ImageCLEF 2026 Deepfake Detection Task in the image modality. Each metric corresponds to binary classification accuracy on a specific subset of the evaluation data, where higher values indicate better performance.

Our detector achieved its highest accuracy on the Baseline Deepfakes subset, with a score of **0.7382**. This indicates that the ensemble was effective at detecting organizer-generated deepfakes produced using baseline or reference manipulation methods. The system also performed reasonably on participant-generated deepfakes, achieving **0.6540** on the unweighted subset and **0.6022** when weighted by the realism score of the generated samples.

Performance on real images was lower, with **0.6026** accuracy on Organizers Real Data and **0.5040** on Organizers Real GT Data. Since the verified real subset carries a high weight in the final score, this result suggests that the detector had a tendency to over-predict the fake class, which is consistent with the prediction distribution observed in Table 4. The final weighted image-detection score was **0.5766**.

Table 5

Official leaderboard scores for our ImageCLEF 2026 Deepfake Detection Task image submission.

Evaluation subset	Accuracy
Baseline Deepfakes (Images)	0.7382
Organizers Real Data (Images)	0.6026
Organizers Real GT Data (Images)	0.5040
Participant Deepfake Data (Images)	0.6540
Participant Deepfake Data Weighted (Images)	0.6022
Final Score (Images)	0.5766

4.8. Error Analysis

The score breakdown reveals two complementary failure modes in our system:

False Positives on Real Images. The low accuracy on both real-image subsets (60.3% and 50.4%) indicates that our ensemble over-predicts the fake class. This aligns with the CLIP-ViT detector’s prediction distribution of 66.5% fake and the ensemble’s 57.1% fake. The EfficientNet-B4 branch, which was not fine-tuned on deepfake data and predicted only 17.3% as fake, partially counterbalances this bias through the ensemble, but the 0.7 CLIP-ViT weight means the fake-leaning bias dominates.

Missed Realistic Deepfakes. The drop from 65.4% to 60.2% accuracy when participant deepfakes are weighted by realism demonstrates that our detector is less effective against the most convincing

forgeries. This is particularly relevant given the task’s adversarial design, where Generation participants are incentivized to produce deepfakes that evade Detection systems.

5. Limitations and Future Work

Our submission has several limitations across both the generation and detection subtasks. For generation, the warp-based pipeline achieved strong landmark conformity but remained sensitive to large pose and expression changes. When target landmarks differed substantially from the available source frames, piecewise affine warping introduced interpolation artifacts, especially around the mouth, jawline, and hairline. Although GFPGAN reduced some visual distortions, it occasionally hallucinated fine details such as teeth or eye regions, which could affect identity consistency and introduce detectable artifacts. In addition, our anti-forensic post-processing used fixed parameters for JPEG compression, sensor noise, and blur, which may not be optimal for all images.

For detection, the main limitation was poor calibration on real images. The ensemble tended to over-predict the fake class, leading to weak performance on the hidden real-image subset. This suggests that the CLIP-ViT branch retained dataset-specific biases from its source training data, while the EfficientNet-B4 branch was not fine-tuned for deepfake detection. The fixed decision threshold of 0.5 and the absence of task-specific validation data further limited calibration. Moreover, the detector relied only on spatial-domain cues and did not exploit frequency-domain artifacts, which are often useful for identifying synthetic images.

Future work will focus on improving both robustness and calibration. For generation, dense landmark-conditioned diffusion models or modern face reenactment frameworks could replace geometric warping to better handle extreme poses while preserving identity. Adaptive or adversarial anti-forensic post-processing may also improve evasion without degrading visual quality. For detection, future systems should combine multi-dataset CLIP ensembles, fine-tuned CNN branches, frequency-domain features, and threshold calibration using proxy validation data. Test-time adaptation and multi-scale facial analysis may further improve generalization to unseen real and synthetic image distributions.

6. Conclusion

We presented the MarsadLab team’s image-generation and image-detection systems for the ImageCLEF 2026 Deepfake Detection and Generation task. For generation, our landmark-guided warp-based pipeline successfully produced all 1,002 required images and achieved strong facial-geometry alignment, showing the effectiveness of geometric warping under strict landmark constraints. However, identity preservation and evasion against participant detectors remained challenging.

For detection, our CLIP-ViT and EfficientNet-B4 ensemble achieved a final score of 0.5766. The system performed better on deepfake subsets than on real-image subsets, indicating a tendency to over-predict the fake class. Future work will focus on dense landmark-conditioned generation, improved identity preservation, multi-dataset detector training, frequency-domain features, and better threshold calibration.

Acknowledgment

This work was partially supported by the Qatar Research, Development and Innovation Council (QRDI) under grant NPRP14C-0916-210015. We also acknowledge the use of Claude (Anthropic) and GPT (OpenAI) as AI writing assistants for drafting, revising, and improving the manuscript content.

Declaration on Generative AI

The authors used the following generative-AI tools while preparing this manuscript:

- **Anthropic Claude Opus 4.8 (July 2026)** for paraphrasing, sentence rephrasing, grammar checking, and language polishing.
- **OpenAI ChatGPT GPT-5.5 (July 2026)** for LaTeX formatting, paraphrasing, and rewording.

All AI-assisted content was critically reviewed, revised, and verified by the authors, who accept full responsibility for the integrity and accuracy of the final manuscript.

References

- [1] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, D. Tao, Deepfake generation and detection: A benchmark and survey, *ACM Computing Surveys* 58 (2026) 1–41.
- [2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, *Information fusion* 64 (2020) 131–148.
- [3] B. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, *Calif. L. Rev.* 107 (2019) 1753.
- [4] A. Yermakov, J. Cech, J. Matas, M. Fritz, Deepfake detection that generalizes across benchmarks, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 773–783.
- [5] P. K. Kushwaha, A. Srivastava, D. Bhargava, B. P. Lohani, A. Gupta, Deepfakebench: A comprehensive benchmark of deepfake detection, in: *2025 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, IEEE, 2025, pp. 736–741.
- [6] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot... for now, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.
- [7] 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.
- [8] D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, B. Ionescu, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, Overview of imageclef 2026 deepfake task: Multimodal detection and generation of deepfakes, in: *CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [9] 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).
- [10] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, S. Zafeiriou, Arc2face: A foundation model for id-consistent human faces, in: *European Conference on Computer Vision, Springer, 2024*, pp. 241–261.
- [11] J. Xu, S. Li, J. Wu, M. Xiong, A. Deng, J. Ji, Y. Huang, G. Mu, W. Feng, S. Ding, et al., Identity-preserving-yet-diversified diffusion models for synthetic face recognition, *Advances in Neural Information Processing Systems* 37 (2024) 77777–77798.
- [12] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, K. Lee, Diffface: Diffusion-based face swapping with facial guidance, *Pattern Recognition* 163 (2025) 111451.
- [13] W. Bondurant, A. Sinha, H. Le, S. Das, S. Schuckers, Diffswap++: 3d latent-controlled diffusion for identity-preserving face swapping, *arXiv preprint arXiv:2511.05575* (2025).

- [14] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, G. Tzimiropoulos, Diffusionact: Controllable diffusion autoencoder for one-shot face reenactment, in: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2025, pp. 1–11.
- [15] A. Rochow, M. Schwarz, S. Behnke, Fsr: Facial scene representation transformer for face reenactment from factorized appearance head-pose and facial expression features, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 7716–7726.
- [16] 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.
- [17] E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9459–9468.
- [18] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, First order motion model for image animation, *Advances in neural information processing systems* 32 (2019).
- [19] S. Ha, M. Kersner, B. Kim, S. Seo, D. Kim, Marionette: Few-shot face reenactment preserving identity of unseen targets, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 10893–10900.
- [20] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, B. Wu, Deepfakebench: A comprehensive benchmark of deepfake detection, *arXiv preprint arXiv:2307.01426* (2023).
- [21] 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.
- [22] S. Smeu, E. Oneata, D. Oneata, Declip: Decoding clip representations for deepfake localization, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 149–159.
- [23] G. Cazenavette, A. Sud, T. Leung, B. Usman, Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10759–10769.
- [24] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, H. Li, Dire for diffusion-generated image detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22445–22455.
- [25] H. Zhang, Q. He, X. Bi, W. Li, B. Liu, B. Xiao, Towards universal ai-generated image detection by variational information bottleneck network, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 23828–23837.
- [26] J. Park, A. Owens, Community forensics: Using thousands of generators to train fake image detectors, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 8245–8257.
- [27] Y. Yang, Z. Qian, Y. Zhu, O. Russakovsky, Y. Wu, D³: scaling up deepfake detection by learning from discrepancy, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 23850–23859.
- [28] F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, A bias-free training paradigm for more general ai-generated image detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 18685–18694.
- [29] Y. Wang, Z. Huang, X. Hong, Opensdi: Spotting diffusion-generated images in the open world, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 4291–4301.
- [30] N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, et al., Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024, *arXiv preprint arXiv:2503.02857* (2025).
- [31] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.

- [32] PiecewiseAffineTransforms, <https://github.com/dfdx/PiecewiseAffineTransforms.jl>, 2020. GitHub repository. Accessed: 2026-05-23.
- [33] 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, volume 2024, 2024, pp. 1862–1874.
- [34] 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).
- [35] 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.
- [36] A. Yermakov, J. Cech, J. Matas, Unlocking the hidden potential of clip in generalizable deepfake detection, arXiv preprint arXiv:2503.19683 (2025).
- [37] 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.
- [38] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.