Supplementary Material for: TalkingHeadBench: A Multi-Modal Benchmark & Analysis of Talking-Head DeepFake Detection

Xinqi Xiong[1]*†, Prakrut Patel[1]*, Qingyuan Fan[1]*, Amisha Wadhwa[1]*, Sarathy Selvam[1]*, Xiao Guo[2], Luchao Qi[1], Xiaoming Liu[2], Roni Sengupta[1]†

[1]University of North Carolina at Chapel Hill [2]Michigan State University

{xxiong, prakrut, lqi, ronisen}@cs.unc.edu

{qfan, wamish, sarathy}@unc.edu {guoxia11, liuxm}@msu.edu

*Equal contribution. †Corresponding Authors.

All benchmarks and curated dataset for our TalkingHeadBench are publicly released at https://anaxqx.github.io/talkingheadbench.github.io. TalkingHeadBench is released under a Creative Commons Attribution 4.0 License (CC BY 4.0).

Our supplementary materials are summarized as follows:

- **Details of Assets Used (Section A):** Descriptions, sources, and licenses for source datasets, talking-head generators, and deepfake detectors utilized in our benchmark.
- **Detailed Experimental Setup (Section B):** Dataset splits, data curation procedures, computational resources, and key training hyperparameters.
- **Detailed Per-Generator Artifact Observations (Section C):** Specific types of visual artifacts commonly produced by each generator, noted during our manual curation process.
- **Experimental Results (Section D):** Performance for commercial generator (MAGI-1) and academic generator (Hallo3).
- **Explainability Results (Section E):** Grad-CAM visualizations for benchmarked detectors.

A. Details of Assets Used

A.1. Source Datasets for Generation and Real Videos

TalkingHeadBench leverages publicly available datasets for source images and driving signals (audio/video) for talking-head generation and real videos for detector training and testing.

Source Images for Deepfake Generation

- **FFHQ (Flickr-Faces-HQ) [8]:** This dataset served as the primary source for high-quality, diverse static portrait images used as the base for deepfake generation. It contains 70,000 high-resolution images of diverse human faces with diverse ages, ethnicities, and image backgrounds originally intended for GAN benchmarking. We utilized images from the 1024x1024 resolution subset. The dataset can be accessed through the following link: https://github.com/NVlabs/ffhq-dataset under Creative Commons BY-NC-SA 4.0 license by NVIDIA Corporation.

Source Driving Signals (Audio/Video) for Deepfake Generation

- **CelebV-HQ [19]:** This large-scale video facial attributes dataset was used as the source for dynamic driving signals. It contains a variety of identities, ethnicities, expressions, and poses. Both video segments (for video-driven generators) and extracted audio tracks (for audio-driven generators, converted to .wav format) were utilized. The dataset can be accessed through the following link: https://github.com/CelebV-HQ/CelebV-HQ. The copyright information for this dataset is not explicitly mentioned.

Real Videos for Detector Training/Testing

- **FaceForensics++ (FF++) [10]:** Original, unmanipulated video sequences were included in our set of real videos. We directly downloaded the dataset following the download script they provided, and finalized the number of our YouTube real videos from 1000 to 704 due to the unavailability of some of the videos on YouTube. We then renamed the downloaded audio to match the naming of the YouTube real videos. The renamed audio files are in our Hugging Face datasets (./audio/ff++). The original FF++ dataset can be accessed through the following link: https://github.com/ondyari/

FaceForensics. The copyright information can be found at https://github.com/ondyari/FaceForensics/blob/master/LICENSE.

- **CelebV-HQ [19]:** To diversify the identity in our real videos, we incorporated a distinct set of real videos from CelebV-HQ that were not used for driving signal extraction.
- **Identity Control:** To prevent identity leakage and ensure fair evaluation, rigorous identity separation was maintained. Identities in the real video sets did not overlap with those used for generating deepfakes (either source images or driving signals) or across the train/test splits of the real videos themselves. This was verified using face recognition (InsightFace with ArcFace model on middle frames of videos).

A.2. Talking-Head Generators

The deepfake videos in TalkingHeadBench were generated using the following state-of-the-art talking-head generators. We provide brief descriptions, primary modality, and links to code repositories/licenses where available. Visual examples are in Fig. 1.

Hallo [16]

- **Description:** Hallo is a talking-head generator that was published in 2024. The generator aims to create realistic and temporally consistent talking-head animations from a static portrait image and a driving audio clip. Hallo integrates diffusion-based models, a UNet-based denoiser, temporal alignment techniques, and a reference network. This hierarchical audio-driven method allows for diversity in poses and expressions. This end-to-end approach enhances animation quality, motion diversity, and personalization across different identities.
- **Modality:** Audio-driven.
- **Code Repository:** https://github.com/fudan-generative-vision/hallo
- **License:** MIT License.
- **Key Artifacts Observed:** See Section C.2.

Hallo2 [1]

- **Description:** Hallo2 is an audio-driven talking-head generator that extends Hallo by enabling long-duration and high-resolution video synthesis published in ICLR2025. It incorporates enhanced temporal modeling and spatial fidelity mechanisms to produce stable and expressive animations over extended sequences. Built upon a diffusion-based generation backbone, Hallo2 introduces improved temporal alignment and rendering strategies that support better lip-sync accuracy and identity preservation. Hallo2 is suited for real-world applications such as virtual avatars, video dubbing, and personalized content creation.
- **Modality:** Audio-driven.
- **Code Repository:** https://github.com/fudan-generative-vision/hallo2
- **License:** MIT License.
- **Key Artifacts Observed:** See Section C.3.

Hallo3 [2]

- **Description:** Hallo3 is an audio-driven portrait image animation model that uses a diffusion transformer backbone, with an identity reference network combining a causal 3D VAE plus stacked transformer layers, to generate highly dynamic, realistic video from static portraits. It improves over U-Net based generators by better handling non-frontal perspectives, immersive backgrounds, and motion dynamics, while preserving identity consistency even across viewpoint changes.
- **Modality:** Audio-driven.
- **Code Repository:** https://github.com/fudan-generative-vision/hallo3
- **License:** MIT License.
- **Key Artifacts Observed:** Generally high quality; specific subtle artifacts, if any, are less pronounced or systematic compared to some open-source academic models.

AniPortrait [15]

- **Description:** AniPortrait is a talking-head generator introduced in 2024. The two-stage approach first extracts 3D representations from audio, converting them to 2D facial landmark sequences. Then, a diffusion model with a motion module transforms these landmarks into temporally coherent talking-head videos. The hierarchical audio-driven pipeline offers precise control over audio-driven facial expressions and head poses while maintaining identity through reference guidance. Its training scheme, which decouples identity encoding from motion dynamics, enables an end-to-end approach that produces animations with accurate lip-sync, detailed expressions, and robust identity preservation. In video-driven mode, AniPortrait directly extracts facial landmarks from a source video to transfer expressions and movements to the reference portrait, utilizing the same reenactment pipeline for consistent temporal guidance and high-quality results.
- **Modality:** Audio-driven or video-driven.
- **Code Repository:** https://github.com/Zejun-Yang/AniPortrait
- **License:** Apache-2.0 license.
- **Key Artifacts Observed:** See Section C.4 for audio-driven artifacts and C.4 for video-driven artifacts.

LivePortrait [5]

- **Description:** LivePortrait is a real-time talking-head generator designed to produce high-quality portrait animations from a single static image introduced in 2024. It employs an implicit keypoint-based architecture combined with

lightweight retargeting and stitching modules to ensure accurate facial motion and head pose transfer. Trained on a large-scale dataset of over 69 million frames, LivePortrait generalizes well across diverse identities and expressions.
- **Modality:** Video-driven.
- **Code Repository:** https://github.com/KwaiVGI/LivePortrait
- **License:** MIT License.
- **Key Artifacts Observed:** See Section C.1.

EMOPortraits [4]
- **Description:** EMOPortraits is a one-shot talking-head generator introduced in CVPR2024. It employs a two-stage training process, with an optional audio-driven phase for video generation from a single image and audio input. The model selects two random frames of the source and driver at each step, adapts the driver frame's motion and expressions onto the source frame to generate the final image. It encodes a source image into a 3D latent feature and identity descriptor, while a motion module extracts pose/expression codes from a driver, resulting in realistic deepfakes under extreme and asymmetric facial expressions.
- **Modality:** Video-driven.
- **Code Repository:** https://github.com/neeek2303/EMOPortraits
- **License:** Apache-2.0 license.
- **Key Artifacts Observed:** See Section C.5.

MAGI-1 [11] (Commercial)
- **Description:** MAGI-1 is a talking-head generator published in 2025. This open-source generator produces realistic, high-quality, temporally consistent videos from text or image prompts. Built on a diffusion transformer architecture, MAGI-1 generates fixed-length videos, enabling real-time streaming and seamless video continuation. With support for large-scale model sizes and long context lengths, it is well-suited for a wide range of creative and generative video applications.
- **Modality:** Primarily Text-to-Video, can optionally add a reference image.
- **Code Repository:** https://github.com/SandAI-org/MAGI-1
- **License:** Apache-2.0 license.
- **Key Artifacts Observed:** Generally high quality; specific subtle artifacts, if any, are less pronounced or systematic compared to some open-source academic models. Focus is often on overall scene coherence rather than just facial animation.

A.3. DeepFake Detectors

The following SOTA deepfake detection models were benchmarked on TalkingHeadBench. Brief descriptions, primary input modality, and links to code repositories/licenses are provided.

CADDM [3]
- **Description:** CADDM utilizes an Artifact Detection Module designed to focus on local regions of images. This module employs multiscale anchors to detect and classify artifact areas, aiming to mitigate the influence of identity information in deepfake detection, making it generalizable across different datasets. It showed great improvement in both accuracy and robustness when evaluating FF++. We directly deployed the model from the official GitHub repository to our cluster without modifying any hyperparameter or model structure.
- **Modality:** Image-based (CNN).
- **Code Repository:** https://github.com/megvii-research/CADDM
- **License:** Apache-2.0 license.

TALL [17]
- **Description:** TALL introduces a temporal-attentive localization and learning framework designed to exploit temporal inconsistencies in deepfake videos. By leveraging a dual-stream architecture that models both short-term and long-term temporal dependencies, TALL isolates manipulated frames and emphasizes temporal transitions typically ignored by frame-based detectors. A temporal attention mechanism further enhances frame-level representations by focusing on abrupt motion irregularities introduced by forgery generation processes. Moreover, the authors propose TALL-Swin, which integrates the TALL strategy with the Swin Transformer architecture. This combination leverages the Swin Transformer's hierarchical feature representation and shifted window mechanism to effectively model both local and global dependencies within the thumbnail layout. Given there is no available model released on their original GitHub repository, we pretrained the model using FF++ based on the implementation version in DeepfakeBench [18].
- **Modality:** Video-based (Transformer).
- **Code Repository:** https://github.com/rainy-xu/TALL4Deepfake
- **License:** MIT license.

LipFD [9]
- **Description:** LipFD targets lip-sync inconsistency by focusing on the alignment between lip motion and audio speech content. The model extracts fine-grained spatio-temporal features of mouth regions and correlates them with phoneme-level audio embeddings to detect subtle mismatches indicative of forgery. By isolating this cross-modal inconsistency, LipFD distinguishes itself from generic detectors that rely primarily on visual features.

This focused approach leads to robust detection of audio-driven deepfakes, especially under real-world conditions. We directly reproduced the results from official GitHub repository with minimal adjustments on learning rate and learning rate decay.
- **Modality:** Audio-Visual (Transformer).
- **Code Repository:** https://github.com/AaronComo/LipFD
- **License:** N/A

DeepFake-Adapter [12]
- **Description:** DeepFake-Adapter presents a universal adaptation framework for deepfake detection by leveraging modality specific adapters integrated into a unified transformer backbone. These adapters specialize in capturing subtle, forgery-specific artifacts across diverse data modalities (e.g., RGB, Depth, Frequency), enabling cross-modal generalization without retraining. We directly reproduced the results from official GitHub repository with minimum adjustments to train using our custom dataset.
- **Modality:** Image-based (Transformer).
- **Code Repository:** https://github.com/rshaojimmy/DeepFake-Adapter
- **License:** N/A

AltFreezing [14]
- **Description:** A video-based deepfake detector that leverages a spatiotemporal model. It employs a training strategy for 3D ConvNet video detectors that alternately freezes spatial- and temporal-related parameter groups to force learning of both artifact types, yielding stronger out-of-distribution generalization on face forgery videos.
- **Modality:** Video-based (3D ConvNet / spatiotemporal).
- **Code Repository:** https://github.com/ZhendongWang6/AltFreezing
- **License:** MIT license.

MM-Det [13]
- **Description:** A diffusion deepfake video detector that builds a Multi-Modal Forgery Representation (MMFR) using a large multi-modal model (LLaVA) and fuses it with a spatiotemporal backbone enhanced by In-and-Across Frame Attention (IAFA). Introduces the DVF dataset and reports SOTA on diffusion-generated videos.
- **Modality:** Video-based (Diffusion).
- **Code Repository:** https://github.com/SparkleXFantasy/MM-Det
- **License:** Apache-2.0 license.

HiFi-Net [6]
- **Description:** An image forgery detection & localization framework with a hierarchical fine-grained formulation: multi-branch feature extractor plus dedicated localization and classification heads to capture subtle artifacts across manipulation types.
- **Modality:** Image-based (detection + localization).
- **Code Repository:** https://github.com/CHELSEA234/HiFi_IFDL
- **License:** MIT license.

B. Detailed Experimental Setup

B.1. Dataset Generation, Splits, and Curation Details

Deepfake Video Generation Approximately 500-600 videos were initially generated for each of the six open-source academic models. This was achieved by scripting random pairings of source images from FFHQ [8] with driving signals (audio or video) from distinct splits of CelebV-HQ [19].

Data Curation and Quality Control A rigorous manual curation process was undertaken to ensure the quality and challenge of the benchmark. This involved:
- Reviewing each generated video for visual fidelity and realism.
- Removing videos with obvious generation failures, extreme distortions, or artifacts that make them trivially identifiable as fake (unless such artifacts are characteristic and subtle).
- Ensuring that the remaining fakes posed a reasonable challenge to detection models.

This process resulted in the final dataset sizes reported in Tab.4 of the main paper, with approximately 60-65% of initially generated videos being discarded.

Train/Test Splits and Real Data Integration
- **Identity Separation:** Strict identity separation was enforced for all data. Source identities from FFHQ and driving signal identities from CelebV-HQ used for the training set of deepfakes were disjoint from those used for the test set.
- **Real Videos:** Real videos from FF++ and CelebV-HQ were also split into training and testing sets, maintaining identity separation. The number of real videos was balanced to be approximately 1:1 with fake videos in the training splits for each protocol.
- **Validation Sets:** For model validation during training, a small subset of identities from the training pool (both real and fake) was held out. For Protocol 1, this involved 50 real and 50 fake videos. For Protocols 2 and 3, it was 50 real and 50 fake videos per generator.

B.2. Computational Resources

The generation of TalkingHeadBench and the benchmarking of detection models were conducted using NVIDIA RTX A4500 GPUs. Deepfake generation times varied significantly depending on the generator model and the number of GPUs used, as videos were generated in parallel across multiple GPUs. On average, producing 500 videos took several hours to up to a day. For detector training, most state-of-the-art models completed training within 4 hours on a single GPU, with the exception of one model that required approximately an hour per epoch—making its total training time dependent on the specific protocol used.

B.3. Hyperparameters and Training Details for Detectors

For all benchmarked detectors, we adhered closely to the training configurations and hyperparameters proposed in their original publications and official codebases.

C. Detailed Per-Generator Artifact Observations

This section details characteristic visual artifacts observed for each generator during the manual data curation phase. These insights informed our curation and highlight the unique challenges posed by different generation techniques.

C.1. LivePortrait [5]

Common artifacts observed in generations from the LivePortrait model included:

- **Non-Physical Head Kinematics and Scaling:** Driving videos featuring substantial translational or lateral head motion often induced unnatural scaling (e.g., apparent growth or shrinkage) or other non-physical movements in the synthesized head, indicative of inconsistent head movement/pose and face warping/distortion.
- **Sensitivity to Source Image Composition:** The generation process exhibited heightened susceptibility to artifacts when source images contained non-adult subjects or multiple individuals. In scenarios with multiple people, artifacts manifested as face warping/distortion, erroneous animation of partially occluded or out-of-frame figures, and visible errors related to internal bounding box estimations, leading to significant spatial inconsistencies.
- **Semantic Misinterpretation and Occlusion Errors:** Objects proximal to the head in the source image were sometimes erroneously segmented as facial features, resulting in their co-movement with the head, a form of occlusion error or texture anomaly.
- **Lip Synchronization Deficiencies with Extreme Expressions:** Source images depicting pronounced oral expressions (e.g., open mouths, broad smiles, compressed lips) occasionally resulted in static labial regions or aberrant lip articulation, indicating lip sync errors or failures in modeling extreme facial deformations.

C.2. Hallo [16]

Common artifacts observed in generations from the Hallo model included:

- **Lip Synchronization Discrepancies:** Generated videos frequently exhibited a temporal mismatch between the audible speech and the synthesized lip movements, a common form of lip sync error.
- **Hair Rendering Artifacts:** Portions of the synthesized hair often displayed unnatural stasis, appeared to adhere to the background, or were rendered with low fidelity (e.g., hair artifacts), particularly noticeable during dynamic head movements. This can be considered a type of temporal inconsistency or texture anomaly.
- **Occlusion Handling Deficiencies:** The model demonstrated improper rendering when facial regions were expected to be occluded by elements such as eyeglasses, hair, or hands (if present in the source), leading to occlusion errors.
- **Background Distortion:** The background region immediately surrounding the synthesized head was prone to background warping/distortion correlating with the head movement.

C.3. Hallo2 [1]

Generations from Hallo2 exhibited artifacts similar to its predecessor, Hallo, albeit with some distinct manifestations:

- **Aberrant or Absent Lip Articulation:** Instances of absent lip movement despite audible speech, or significant asynchrony between audio and visual lip cues, were noted, representing pronounced lip sync errors and temporal inconsistencies.
- **Static Peripheral Hair Artifacts:** Hair situated outside the primary head bounding box frequently remained static during head motion, creating a visually jarring "detached" effect—a specific type of hair artifact and temporal inconsistency.
- **Object-Induced Spatial Distortion:** The presence of objects proximate to or intersecting with the head's bounding box in the source image often precipitated localized face warping/distortion or spatial inconsistencies in the output.
- **Background-Correlated Visual Anomalies:** Unsystematic visual artifacts, such as unpredictable patterns or noise, occasionally manifested, appearing to be correlated with complex textures or patterns within the source image's background, a form of spatial inconsistency or texture anomaly.

C.4. AniPortrait [15]

Audio-driven Analysis of audio-driven outputs from AniPortrait revealed the following:

- **Variable Perceptual Realism:** The model demonstrated capacity for producing naturalistic results, particularly notable for an exclusively audio-driven synthesis approach.
- **Synchronization and Kinematic Irregularities:** The system was prone to occasional lip sync errors and exhibited unnatural head kinematics, including sudden accelerations, oscillatory movements, or an overly rigid, static posture, all categorized under inconsistent head movement/pose.
- **Dental Structure Anomalies:** Synthesized dentition sometimes presented with unrealistic characteristics, such as supernumerary or atypically arranged rows of teeth (teeth anomalies).
- **Pupillary Artifacts with Eyewear:** Source images featuring subjects with eyeglasses occasionally led to anomalous pupillary dilation patterns, such as variable frequencies of dilation (pupil anomalies).
- **Hand-Induced Artifacts:** The presence of hands within the source image frame was a primary trigger for various occlusion errors and spatial inconsistencies.

Video-driven Video-driven synthesis using AniPortrait was characterized by:

- **Microphone Occlusion Handling:** The model displayed an unusual proficiency in generating plausible outputs when the driving video featured a subject with a microphone in close proximity to the mouth, suggesting effective handling of this specific occlusion scenario.
- **Catastrophic "Head Explosion" Artifacts:** A frequent failure mode involved severe face warping/distortion, where the synthesized head appeared to disintegrate or become overlaid with disparate visual elements (often misidentified hands or other body parts) from the driving video, a critical spatial inconsistency.
- **Pervasive Background Instability:** Background warping/distortion was commonly found even in outputs that were otherwise subjectively assessed as high quality.
- **Motion Tracking Fidelity versus Distortion Thresholds:** While the model effectively tracked subtle head movements, more pronounced motions (e.g., swaying) frequently surpassed its stable generation threshold, resulting in severe face warping/distortion that could occupy a significant portion of the video frame, indicating limitations in maintaining inconsistent head movement/pose coherence.
- **Hair Segmentation Artifacts:** In certain instances involving female source images, synthesized hair was subject to abrupt truncation or unnatural rendering.
- **Severe Hand-Related Artifacts:** The model exhibited a notable inability to process hand movements in the driving video, almost invariably leading to prominent and clearly delineated occlusion errors and spatial inconsistencies.
- **High Incidence of Severe Generation Failures:** A substantial proportion of initial generations were unusable due to extreme face warping/distortion, transforming the source image into an unrecognizable amalgam of textures and features, indicative of fundamental failures in identity preservation and coherent image synthesis.

C.5. EMOPortraits [4]

Artifacts observed in EMOPortraits generations included:

- **"Floating Head" Artifacts:** Facial and head modifications were often restricted to the cephalic region, leading to a dissociation from neck and body movements. Significant corporeal motion in the driving video could result in the head appearing detached or "floating," a clear example of inconsistent head movement/pose and spatial inconsistency.
- **Perifacial Blurring and Edge Anomalies:** The model frequently introduced blurring in the regions immediately surrounding the synthesized face, potentially diminishing perceptual realism and constituting an edge anomaly or texture anomaly.
- **Off-Axis Pose Instability:** The system encountered difficulties rendering non-frontal (e.g., three-quarter or full profile) views. Lateral head rotations could precipitate severe face warping/distortion, including apparent shrinkage, color mismatch or degradation into noise resembling video static.
- **Eyewear-Related Artifacts:** The model struggled with subjects wearing eyeglasses, particularly sunglasses. This often resulted in occlusion errors or texture anomalies where eyes were unrealistically rendered as visible through otherwise opaque lenses, or pupil anomalies.
- **Chroma Key-like Color Artifacts:** Generated videos frequently exhibited spurious patches of bright, uniform color (commonly green), reminiscent of poorly executed chroma keying, indicating significant color mismatch or spatial inconsistencies.

D. Additional Experimental Results

D.1. MAGI-1 Commercial Generator Results

We present detection performance on deepfakes generated by the MAGI-1 commercial model, an unseen generator included to test detector robustness against proprietary systems. Since MAGI-1 does not provide audio output, we restrict evaluation to detectors operating on visual input: CADDM, HiFi-Net, TALL, and DeepFake-Adapter (DF-Adapter). Results are reported in Tab. D1.

Overall, **TALL** achieves perfect scores across all metrics (AUC, T1, T0.1), indicating strong robustness to MAGI-1's generation pipeline. **DF-Adapter** also performs near-perfectly, though its performance drops slightly at stricter thresholds (e.g., T0.1 falling to 0.99 on some splits). By contrast, **CADDM** achieves only moderate performance (AUC 0.88–0.97, T1 around 0.3–0.6), reflecting limited general-

ization under MAGI-1. **HiFi-Net** performs worst overall, with low T1 values (< 0.39) and near-zero T0.1 on most splits, suggesting that its learned features fail to transfer to this commercial generator.

These results highlight two key trends. First, detectors with strong multi-generator training performance (TALL, DF-Adapter) transfer well to unseen commercial generators, achieving excellent or near-excellent robustness. Second, detectors such as CADDM and HiFi-Net, which are more fragile under distribution shifts, fail to generalize effectively despite maintaining higher AUC values. This reinforces the broader insight from our benchmark: aggregate metrics such as AUC can obscure vulnerabilities, while threshold-based measures (T1, T0.1) reveal substantial gaps in robustness to emerging generators.

D.2. Hallo3 Academic Generator Results

We evaluate detector robustness on Hallo3, a recent academic talking-head generator. Results are shown in Tab. D2. As with MAGI-1, we restrict evaluation to detectors operating on visual input.

TALL achieves perfect scores across all metrics (AUC, T1, T0.1), demonstrating excellent generalization to Hallo3. **DF-Adapter** also performs nearly perfectly, with T1 and T0.1 values consistently at 0.99–1.00 across all splits. By contrast, **CADDM** shows only moderate performance, with AUC ranging from 0.67 to 0.97 and T1 values spanning 0.24–0.91 depending on the split. **HiFi-Net** performs inconsistently: while it reaches very high scores on EMOPortraits (AUC = 1.00, T1 = 0.99), it collapses on other splits (e.g., T1 = 0.19 on AniAudio, T0.1 = 0.00 on LivePortrait), highlighting its fragility under generator and protocol shifts.

These findings mirror broader trends from our benchmark: detectors such as TALL and DF-Adapter maintain strong robustness on unseen generators, while CADDM and HiFi-Net struggle to generalize reliably. The variability in CADDM and the failure cases of HiFi-Net underscore the importance of evaluating with stricter thresholds, as aggregate metrics like AUC alone would obscure these weaknesses.

E. Additional Explainability Results

The main paper (Section 4.4) provides Grad-CAM visualizations for the TALL detector. This section includes additional Grad-CAM results for TALL and DeepFake-Adapter, the best two benchmarked detectors. These visualizations illustrate the image regions these models focus on, offering insights into their decision-making and potential biases across different generators and protocols.

E.1. Grad-CAM Visualizations for TALL [17]

Figure E1 presents Grad-CAM visualizations of the TALL detector across Protocols 1–3, using one correctly classified high-confidence fake sample per generator per protocol.

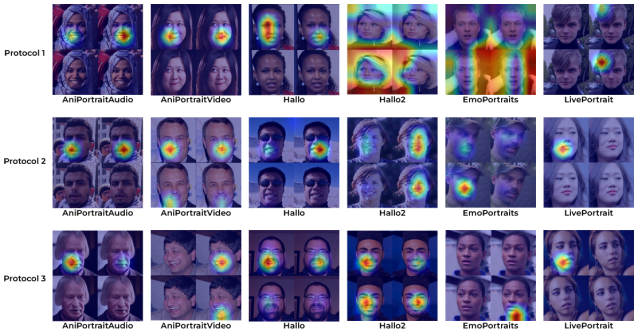

Figure E1. Grad-CAM visualizations of the TALL detector on correctly classified, high-confidence fake samples across Protocols 1–3 (one sample per generator). TALL predominantly focuses on facial features such as the eyes, nose, and mouth, but strategically shifts its attention to background or peripheral regions for certain generators. In Protocol 1, where the model has seen all generators during training, it focuses on background cues for Hallo2 and EMOPortraits, suggesting that these regions carry distinctive generative artifacts. In Protocol 3, the detector localizes to the neck region for EMOPortraits, while maintaining facial focus elsewhere. These examples reflect TALL's adaptive attention and its ability to leverage generator-specific cues that contribute to its superior generalization and detection performance.

These examples reflect cases where TALL confidently and accurately identifies manipulations, allowing us to interpret what visual evidence it relies on for detection.

Across most generators and protocols, TALL focuses its attention primarily on the facial region—including key semantic features such as the eyes, nose, and mouth—indicating that it has learned to localize and exploit common deepfake artifacts. Notably, in Protocol 1, where the model is trained on all generators (i.e., no generator shift), TALL relies exclusively on background regions for Hallo2 and EMOPortraits. This behavior contrasts with its facial focus for other generators and suggests that TALL has learned to exploit generator-specific cues in the background—clues that other detectors likely overlook.

In Protocol 3, where both identity and generator differ between train and test, TALL continues to rely on consistent facial features for most generators. However, it attends to the neck region for EMOPortraits, which may again indicate an adaptation to persistent artifacts in that region. Importantly, all examples shown here are correct classifications, underscoring that TALL's attention—whether focused on facial or peripheral features—is meaningfully aligned with its high performance.

These visualizations suggest that TALL does not overfit to a single detection heuristic but instead generalizes flexibly across generators by identifying the most informative regions—be they facial or contextual—for each case. This ability to adaptively shift focus may contribute to its strong generalization under distribution shifts.

Table D1. Supplementary results on MAGI-1 commercial generator.

	AniAudio [15]			AniVideo [15]			Hallo [16]			Hallo2 [1]			EMOPortraits [4]			LivePortrait [5]		
Detector	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1
CADDM [3]	0.88	0.29	0.29	0.95	0.39	0.39	0.91	0.53	0.53	0.96	0.62	0.62	0.97	0.62	0.62	0.97	0.64	0.64
HiFi-Net [6]	0.66	0.02	0.00	0.89	0.39	0.23	0.85	0.04	0.02	0.73	0.02	0.00	0.86	0.06	0.04	0.87	0.09	0.02
TALL [17]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DF-Adapter [12]	1.00	1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Table D2. Supplementary results on Hallo3 academic generator.

	AniAudio [15]			AniVideo [15]			Hallo [16]			Hallo2 [1]			EMOPortraits [4]			LivePortrait [5]		
Detector	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1	AUC	T1	T0.1
CADDM [3]	0.67	0.24	0.24	0.70	0.40	0.40	0.93	0.76	0.76	0.82	0.64	0.64	0.97	0.91	0.91	0.93	0.69	0.69
HiFi-Net [6]	0.84	0.19	0.18	0.88	0.30	0.27	0.91	0.24	0.20	0.65	0.24	0.12	1.00	0.99	0.99	0.87	0.02	0.00
TALL [17]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DF-Adapter [12]	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99

E.2. Grad-CAM Visualizations for DeepFake-Adapter [12]

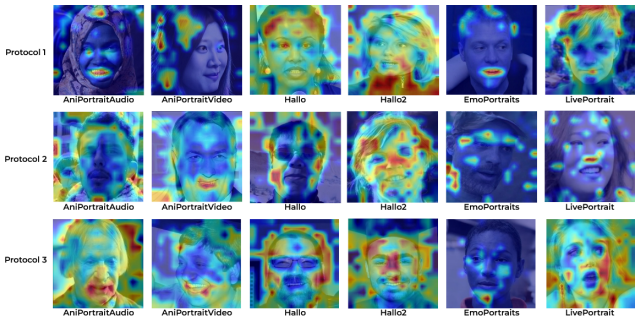

Figure E2. Grad-CAM visualizations of DeepFake-Adapter across Protocols 1–3 (same samples as used for TALL). Compared to TALL, DF-Adapter shows broader and less localized attention maps, often covering both facial and background regions with reduced spatial focus. While it identifies meaningful areas in some cases (e.g., the mouth region in EMOPortraits under Protocol 1), attention weakens in Protocols 2 and 3, especially for EMOPortraits, correlating with its performance drop. The model appears to rely on a mixture of weak signals across the image rather than generator-specific artifact patterns, reflecting its less consistent generalization under distribution shifts.

Figure E2 presents Grad-CAM results for the DeepFake-Adapter model across Protocols 1–3 using the same sample images previously analyzed for TALL. Compared to TALL, whose attention maps are generally well-localized around facial or generator-specific regions, DF-Adapter displays broader and more diffuse attention, often spanning both the face and background without a clearly focused region of interest.

In many cases, DF-Adapter does attend to relevant facial areas—particularly the mouth and eyes—but the heatmaps suggest less spatial precision. This pattern is observed across most generators and protocols, indicating that DF-Adapter may rely on a combination of weak signals distributed across the image rather than distinct artifact cues. For example, in Protocol 1, the attention map for EMOPortraits is focused around the mouth, a region commonly manipulated in talking-head deepfakes. However, in Protocols 2 and 3, the attention on EMOPortraits degrades significantly, with minimal relevant focused regions, potentially explaining the model's poor generalization to this generator under unseen conditions.

The results also highlight DF-Adapter's less adaptive generalization behavior: while it sometimes leverages background information (as TALL does for Hallo2 and EMOPortraits), it lacks the targeted selectivity shown by TALL. This likely contributes to its performance drop under Protocols 2 and 3, where subtle artifacts become harder to detect without robust, focused visual strategies.

F. Demographic Bias Audit

To assess the fairness of the benchmarked detectors and understand biases inherited from the source datasets (FFHQ, CelebV-HQ), we conducted a demographic audit on our test set using classifications from the FairFace model [7]. We analyzed detector performance across three axes: race, gender, and age. Tab F3 summarized below indicate minimal bias along gender lines but reveal notable performance disparities across racial and age subgroups. This analysis highlights an important area for future work in both dataset creation and algorithmic fairness for deepfake detection.

Table F3. Detector performance by demographics. Disparities appear across race and age, while gender differences are minimal.

(a) Race

Group	Accuracy
Asian	88.89%
African American	81.58%
Indian	87.23%
White	87.55%

(b) Gender

Group	Accuracy
Female	87.77%
Male	86.98%

(c) Age Group

Age Group	Accuracy
0–2	95.00%
3–9	93.88%
10–19	90.91%
20–29	87.90%
30–39	84.48%
40–49	82.56%
50–59	85.37%
60–69	94.12%
70+	80.00%

References

[1] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 2, 5, 8

[2] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. *arXiv preprint arXiv:2412.00733*, 2024. 2

[3] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023. 3, 8

[4] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 3, 6, 8

[5] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2, 5, 8

[6] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 4, 8

[7] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 8

[8] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks, 2019. 1, 4

[9] Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. *Advances in Neural Information Processing Systems*, 37:91131–91155, 2024. 3

[10] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1

[11] Sand-AI. Magi-1: Autoregressive video generation at scale, 2025. 3

[12] Rui Shao, Tianxing Wu, Liqiang Nie, and Ziwei Liu. Deepfake-adapter: Dual-level adapter for deepfake detection. *International Journal of Computer Vision*, pages 1–16, 2025. 4, 8

[13] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection, 2025. 4

[14] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4129–4138, 2023. 4

[15] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 2, 5, 8

[16] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 2, 5, 8

[17] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 3, 7, 8

[18] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: a comprehensive benchmark of deepfake detection. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 4534–4565, Red Hook, NY, USA, 2023. Curran Associates Inc. 3

[19] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 1, 2, 4