# X-AVDT: Audio-Visual Cross-Attention for Robust Deepfake Detection

## Supplementary Material

In this supplementary material, we provide expanded details on the proposed model and the data, an extended ablation study, a class separability analysis, and additional visualizations:

- In Section A, we present additional technical details of our training setup, including the inversion procedure, attention feature extraction and the model architecture. We also describe how the baselines were trained, and detail the human evaluation.
- In Section B, we report the results of experiments on broader deepfake benchmarks, and provide comparative analyses with representative audio-visual baselines.
- In Section C, we (i) report the results of an extended ablation study that compares inversion conditions (audio-driven, text-driven, and withou inversion), (ii) provide detailed results under perturbation attacks, and (iii) analyze attention types and diffusion timesteps.
- In Section D, we conduct a class separability analysis using Fisher SNR and LDA margin to quantify the discriminability of the learned representations. We also present extended cross-attention robustness analyses, along with attention map visualizations that support these findings.
- In Section E, we describe how the MMDF training and evaluation data were obtained and present qualitative examples, including our model's input representations and sample dataset visualizations.
- In Section F, we discuss the limitations of our system.

# A. Additional Experimental Details

## A.1. Implementation Details of X-AVDT

### A.1.1. Input Representation

The full procedure of input representation extraction is summarized in Algorithm 1. We follow Hallo [90] with a paired ReferenceNet [35] to encode identity features from the source portrait frame. During DDIM inversion, the denoising U-Net reads these features via cross-attention. For the cross-attention feature in our detector, we sample at an early diffusion step, setting $t^\star = 24$ out of a 1000 step schedule during inversion. We adopt Hallo's hierarchical audio-visual cross-attention mechanism to handle regional masking. We compute lip, expression, and pose masks, apply them as element-wise gates to the cross-attention features and then aggregate the gated features with learned weights. We operate clip-wise on non-overlapping 16 frame segments. If the video length is not divisible by 16, we repeat the last frame to pad to the nearest multiple before feature extraction, and concatenate the per-clip outputs along time. This extraction pipeline is applied identically across all datasets for both training and evaluation.

**Algorithm 1** Audio-driven inversion & reconstruction.

**Input:** Video $x$, Reference frame $x_{\text{ref}}$, Audio $c$,
Masks $\mathcal{M} = \{\mathcal{M}_{\text{full}}, \mathcal{M}_{\text{face}}, \mathcal{M}_{\text{lip}}\}$
**Output:** Video composite $\boldsymbol{\phi} = [x, D(\hat{z}_T), D(\hat{z}_0), r]$,
AV cross-attention feature $\boldsymbol{\psi} = \text{CrossAttn}(H(t),\, c)$
**1. Encode.** $z_0 \leftarrow \text{VAE}_{\text{enc}}(x), \quad e_a \leftarrow \text{Audio}_{\text{enc}}(c)$
**2. Reference pass.** $\text{RefFeat} \leftarrow \text{ReferenceNet}(x_{\text{ref}})$
**3. DDIM Inversion.** Run the inverse scheduler $z_0 \rightarrow z_T$.
**for** $t \in T$ *(fine → coarse)* **do**
  $(\hat{\epsilon}_t,\, \boldsymbol{\psi}_t) \leftarrow \text{UNetFwd}(z_t,\, e_a,\, \mathcal{M};\, \text{RefFeat})$
  $z_{t+1} \leftarrow \text{DDIMInverseScheduler}(z_t, \hat{\epsilon}_t)$
  **if** $t = t^\star$ **then**
    $\tilde{\boldsymbol{\psi}} \leftarrow \text{HeadProj}(\psi_t)$
    $\boldsymbol{\psi} \leftarrow \sum_{k \in \{\text{full,face,lip}\}} w_k \, (\tilde{\boldsymbol{\psi}} \odot \mathcal{M}_k)$
  **end**
**end for**
**4. DDIM Reconstruction.** Run the forward scheduler $z_T \rightarrow z_0$.
**for** $t \in T$ *(coarse → fine)* **do**
  $\hat{\epsilon}_t \leftarrow \text{UNetFwd}(\tilde{z}_t,\, e_a,\, \mathcal{M};\, \text{RefFeat})$
  $\tilde{z}_{t-1} \leftarrow \text{DDIMScheduler}(\tilde{z}_t,\, \hat{\epsilon}_t)$
**end for**
$\hat{z}_0 \leftarrow \tilde{z}_0, \;\; \hat{x} \leftarrow D(\hat{z}_0), \;\; u \leftarrow D(\hat{z}_T), \;\; r \leftarrow |x - \hat{x}|,$
$\boldsymbol{\phi} \leftarrow [x,\, u,\, \hat{x},\, r]$
**Return** $(\boldsymbol{\phi},\, \boldsymbol{\psi})$

### A.1.2. Conditioning

We use wav2vec 2.0 [6] as the audio feature encoder to condition our videos. To capture rich semantics information across different audio layers, we concatenate the audio embeddings from the last 12 layers of wav2vec 2.0 network. Given the sequential nature of audio, we aggregate a 5-frame local context (t−2…t+2) for each video frame before projection.

### A.1.3. Training

To fuse the video composite $\boldsymbol{\phi}$ and AV cross-attention feature $\boldsymbol{\psi}$ during training, we proceed as follows. We concatenate $\mathbf{v}'$ and $\mathbf{a}'$ along the channel dimension and apply a $1\times1$ convolution, reducing the channels from 2048 to 1024 to obtain $p_i$. We add fixed 2D positional encodings to $p_i$ and apply an 8-head self-attention over the $HW$ tokens, with LayerNorm and a residual connection. We feed the self-attention outputs into three 3D ResNeXt [89] layers, followed by global average pooling, which yields $g_i \in \mathbb{R}^{1024}$. We train for 2 epochs by default, as our inputs are structured internal representations extracted from a pretrained diffusion model, enabling faster convergence than raw RGB. Table A.1.3 reports an ablation result showing that performance quickly converges after a few epochs.

| Epochs | 1 | 2 (Ours) | 5 | 10 | 20 |
|---|---|---|---|---|---|
| AUROC | 93.24 | **95.29** | 95.01 | 95.19 | 95.13 |

Table A.1.3. **Effect of training epochs on X-AVDT.**

## A.2. Details of Baseline Detectors

### A.2.1. LipForensics [31]

LipForensics is a video-only deepfake detector that operates on mouth crops, targeting the lip region and modeling temporal inconsistencies in mouth movements to identify manipulation-specific irregularities. We evaluated LipForensics using the official pretrained model that has been trained on FaceForensics++. In addition, we did not retrain it because the training code is not available.

### A.2.2. RealForensics [32]

RealForensics uses audio-visual pretraining, in which audio and visuals exclusively from real samples are used to learn representations that help a classifier discriminate between real and fake videos. We evaluated RealForensics using the official pretrained model that has been trained on FaceForensics++, and we also retrained it on MMDF using the same hyperparameters.

### A.2.3. AVAD [24]

AVAD first pretrains an audio–visual synchronization model following Chen *et al.* [14] to learn temporal alignment between speech and mouth motion. They then use the inferred features to train an anomaly detector, producing a fully unsupervised multi-modal deepfake detector. As an unsupervised method, AVAD is not trained with labels or fake examples. We evaluated AVAD using the official pretrained model that was trained on LRS [79] and did not retrain it because the training code is not available.

### A.2.4. FACTOR [70]

FACTOR is a training-free deepfake detector that frames detection as fact checking. It uses audio-visual encoders to extract modality-specific features and computes a truth score (cosine similarity) that quantifies the consistency between observed audio-visual evidence and an asserted attribute. FACTOR operates in a zero-shot, label-free setting and does not use fake examples. We evaluated FACTOR using the official implementation and pretrained feature extractors.

### A.2.5. LipFD [54]

LipFD targets lip-syncing forgeries by enforcing audio-visual temporal consistency between lip movements and audio signals. The method operates on mouth crops and combines a global video branch with a lip-region branch in a dual-head design. We evaluated LipFD using the model pretrained on Lip Reading Sentences 3 (LRS3) [5], FaceForensics++, and the Deepfake Detection Challenge Dataset (DFDC) [22]. For retraining on MMDF, we trained LipFD for 25 epochs and otherwise keep the original hyperparameters.

| FakeAVCeleb [39] | | | FaceForensics++ [72] | | |
|---|---|---|---|---|---|
| FSGAN | FaceSwap | Wav2Lip | Deepfakes | FaceSwap | Face2Face |
| 99.73 | 99.79 | 99.92 | 99.62 | 99.24 | 99.63 |

Table B.1. **In-domain AUROC on the benchmark dataset.** Cross-manipulation generalization is evaluated by training on two manipulation methods and testing on the remaining one (e.g., train on the first two columns and test on the third columns).

### A.2.6. AVH-Align [78]

AVH-Align addresses dataset shortcuts such as leading silence by training only on real data and learning a frame-level audio-video alignment score from AV-HuBERT features [76]. The training is self-supervised and label-free on real pairs, with no fake examples used. We evaluated AVH-Align using the official pretrained model that has been trained on FakeAVCeleb and AV-Deepfake1M. In addition, we retrained it on MMDF using the same hyperparameters.

## A.3. Human Evaluation

We conducted a human evaluation study to assess deepfake detection accuracy and to quantify the realism of manipulated videos in MMDF, comparing results against FaceForensics++ and FakeAVCeleb. For each clip, participants responded to two following questions: (i) *"Is the video real or fake?"* (binary choice), and (ii) *"What did you focus on when deciding whether the video was real or fake?"*. For question (ii), 80% of comments cited audio-visual synchronization as the primary cue, while the remainder pointed to background artifacts, expression dynamics, and intraoral details, etc. We used 80 videos (60 from the MMDF dataset, 10 from the FakeAVCeleb, and 10 from FaceForensics++), with 24 participants providing responses. We aggregated answers to compute Human Evaluation (HE) accuracy and Human False Acceptance Rate (HFAR), with both metrics derive from the same study.

# B. Additional Experiments

## B.1. In-domain Evaluation

Table B.1 summarizes the in-domain performance of X-AVDT on each benchmark dataset (FakeAVCeleb [39] and FaceForensics++ [72]), where the model is trained and tested on the same dataset ($\sim$ *Official-pretrained*). X-AVDT achieved high AUROC across manipulation types (higher than the scores obtained by the prior methods presented in Table 5). Note that Table 5 reports the results for cross-dataset robustness (MMDF $\rightarrow$ benchmark); while the performances of many *MMDF-trained* baselines dropped due to domain mismatch, X-AVDT remained strong.

## B.2. Additional Experiments

In addition to Table 4 in the main paper, we report additional quantitative results to further evaluate the generalization of SpeechForensics [50]. Table B.2.1 compares X-AVDT with SpeechForensics using representative audio-visual baselines on the MMDF dataset. Although

| Model | AniPortrait [87] | | | MegActor-Σ [91] | | | HunyuanAvatar [15] | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUROC | AP | Acc@EER | AUROC | AP | Acc@EER | AUROC | AP | Acc@EER | AUROC | AP | Acc@EER |
| SpeechForensics [50] | 99.99 | 99.99 | 99.88 | 98.69 | 98.82 | 94.46 | 92.12 | 91.98 | 82.90 | 96.93 | 96.93 | 92.41 |
| **X-AVDT (Ours)** | **99.10** | **98.89** | **96.54** | **90.17** | **88.05** | **83.11** | **97.79** | **97.44** | **97.69** | **95.29** | **94.03** | **91.15** |

Table B.2.1. **Additional quantitative comparison results on the MMDF dataset.**

| *Dataset* | AUROC | AP | Acc@EER | Acc |
|---|---|---|---|---|
| DeepSpeak v1.0 [8] | 94.29 | 95.06 | 95.39 | 94.94 |
| KoDF [41] | 93.07 | 92.88 | 91.13 | 91.87 |
| Deepfake-Eval2024 [12] | 75.02 | 72.73 | 71.68 | 71.36 |

Table B.2.2. **Additional quantitative results for X-AVDT.**

X-AVDT performed worse than SpeechForensics on AniPortrait and MegActor-Σ, it yielded a clear improvement on HunyuanAvatar, where SpeechForensics attained comparatively lower scores. Overall, the results indicate complementary strengths of the two methods across generators and suggest that generator-internal audio-visual consistency cues can be particularly helpful in challenging settings such as HunyuanAvatar, whose high-fidelity, temporally coherent, audio-driven synthesis can suppress overt artifact-based cues. Table B.2.2 summarizes results on additional in-the-wild deepfake benchmarks, such as DeepSpeak v1.0 [8], KoDF [41], and Deepfake-Eval2024 [28]. X-AVDT maintained high performance on DeepSpeak v1.0 and KoDF, but its performance dropped on Deepfake-Eval2024, due to a challenging domain shift in content and compression conditions. Nevertheless, the method remained well above chance across all three datasets, indicating non-trivial generalization in settings with MMDF.

# C. Additional Ablation Study

## C.1. Inversion Condition

In Table C.1, we compare three input settings for the video composite $\phi$. The settings are: text-driven conditioning; original-frame only, which uses only the original RGB frames without the decoded latent DDIM noise map $D(\hat{z}_T)$, the reconstruction $D(\hat{z}_0)$, the residual $r = |x - D(\hat{z}_0)|$, or attention features; and audio-driven conditioning (Ours). For text-driven conditioning, we use the BLIP-2, OPT-2.7b [44] model to caption frames before inversion. The text-conditioned inversion is based on Stable Diffusion 1.5 [71], the same backbone as Hallo. This ablation represents the core designing principle of X-AVDT leveraging internal features of large generative models, specifically audio-visual cross-attention features for deepfake detection. The audio-driven setting consistently yielded the strongest results, validating that audio conditioned cross-attention offers richer, temporally aligned cues than text conditioning. Such alignment is particularly important for facial-editing videos that hinge on subtle mouth and expression edits.

## C.2. Robustness of Perturbation Attack

We conducted an additional ablation study to evaluate the robustness of X-AVDT exploiting audio–visual alignment signals against diverse image perturbations. We assessed performance under five scenarios, with severity 0 denoting the unmodified original. All experiments were run on a subset of MMDF. The baselines (LipForensics [31], RealForensics [32], and AVH-Align [78]) were evaluated with their official pretrained checkpoints.

- **JPEG Compression**: Lossy re-encoding is applied with quality levels of 90, 70, 50, and 30. Lower quality yields stronger high-frequency suppression.
- **Blur**: Gaussian blur with radius of 0.5, 1.0, 2.0, and 3.0, modeling defocus and motion smoothing.
- **Noise**: Additive Gaussian noise with standard deviation (pixel scale) of 5, 10, 20, and 35.
- **Resizing**: Downscale and then upscale using bilinear interpolation with percentage of 75%, 60%, 50%, and 40%. The frame is reduced and then upsampled back to the original size, simulating resolution loss.
- **Frame Drop**: Randomly remove frames with probabilities of 0.05, 0.10, 0.20, and 0.30, creating temporal discontinuities. We do not duplicate frames in this setting.

These experiments were conducted with the same dataset used in Table 4 and Table 5. The quantitative results for AP (%) and Acc@EER (%) are presented in Figure C.2.1. As shown in the results, X-AVDT caused only minor performance degradation across various perturbation methods, while consistently surpassed competing baselines [31, 32, 78], demonstrating strong robustness.

**Audio Perturbation Attack.** We evaluates the robustness of X-AVDT under two audio perturbations: Audio Desynchronization and Audio Codec Artifacts. For desynchronization, we apply a temporal offset $\tau \in \{-0.5, +0.5\}$ seconds, where a positive offset delays speech by prefixing $|\tau|$ seconds of silence, while a negative offset advances audio by trimming the first $|\tau|$ seconds. For codec artifacts, we re-encode audio with a bitrate cap $b \in \{8, 32\}$ kbps to introduce compression distortions. As shown in Table C.2.2, compared to the clean setting ($\tau = 0$, original audio), desynchronization induced only modest performance drops (within 2.4-3.2 points across metrics) at $\tau = \pm 0.5$s. Similarly, compared to no re-encoding, codec artifacts caused limited degradation (within 0.5-7.8 points across metrics) under $b \in \{8, 32\}$ kbps. These results suggest that the learned audio-visual consistency cues remain stable under temporal misalignment and compression-induced distortions.

| Method | AniPortrait [87] | | | MegActor-Σ [91] | | | HunyuanAvatar [15] | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUROC | AP | Acc@EER | AUROC | AP | Acc@EER | AUROC | AP | Acc@EER | AUROC | AP | Acc@EER |
| *w/o* Inversion | 73.55 | 67.56 | 69.95 | 63.18 | 63.69 | 59.81 | 41.81 | 46.36 | 43.55 | 62.22 | 61.07 | 56.69 |
| Text-driven | 90.22 | 90.85 | 81.56 | 76.15 | 75.25 | 69.91 | 49.01 | 49.00 | 50.51 | 74.48 | 76.94 | 65.75 |
| **Audio-driven (Ours)** | **96.55** | **98.83** | **90.20** | **89.87** | **89.25** | **83.86** | **97.71** | **97.04** | **90.94** | **94.71** | **95.04** | **88.33** |

Table C.1. **Ablation on video composite $\phi$.** When training, we fix the backbone and vary only the conditioning signal used during inversion. Audio-driven setting ranked first across datasets, while removing inversion cues yielded the weakest composite.

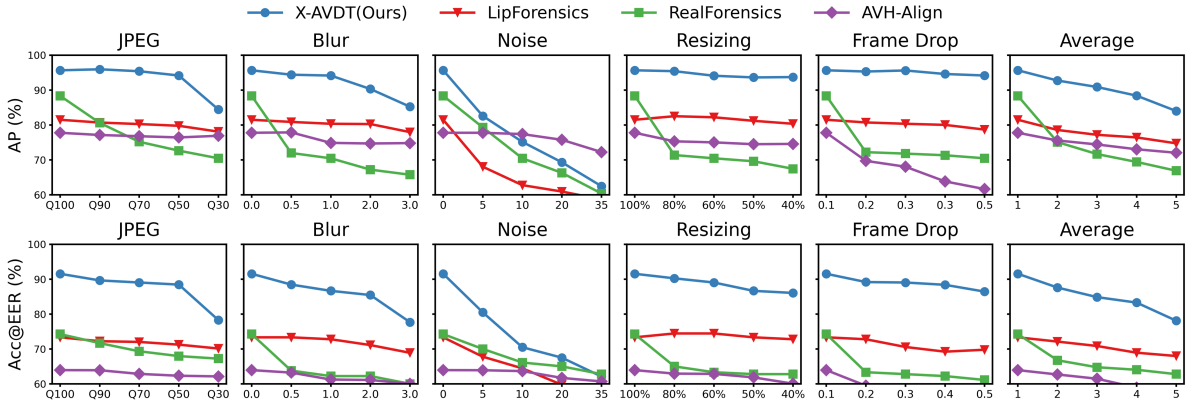



Figure C.2.1. **Robustness to unseen corruptions.** AP (%) is shown in the first row and Acc@EER (%) is shown in the second row across five severity levels. Per corruption scales are shown on the $x$–axes. Average denotes the mean AP in the top row and Acc@EER in the bottom row, respectively, across all corruptions at each severity level.

| Metric | Audio Desynchronizaiton | | | Audio Codec Artifacts | | |
|---|---|---|---|---|---|---|
| | -0.5 sec | 0 | +0.5 sec | 0 | 8k | 32k |
| AUROC | 90.90 | 93.70 | 91.31 | 93.70 | 91.80 | 90.17 |
| AP | 91.10 | 94.30 | 91.74 | 94.30 | 88.80 | 86.51 |
| Acc@EER | 83.40 | 86.40 | 83.97 | 86.40 | 85.90 | 81.56 |

Table C.2.2. **Robustness to unseen audio perturbations.** Performance under audio desynchronization (temporal offsets) and audio codec artifacts (low-bitrate re-encoding) at varying severity levels.

## C.3. Choice of Attention Type and Timestep

We present additional visual examples across different attention types and DDIM inversion timesteps $t$ in Figure C.3. In a diffusion model trained with $T = 1000$ steps, we perform inversion with a 40 step sampling schedule and conducted an ablation study over cross-attention, spatial-attention, and temporal-attention, comparing three representative timesteps at $t \in \{24, 249, 499\}$. Audio-visual cross-attention consistently concentrated on articulators (e.g., lips, jaw), while suppressing the background. Furthermore, as $t$ increased, cross-attention maintained the highest performance at every timestep, outperforming both temporal and spatial attention, indicating that it is the most robust component (see Table 6).

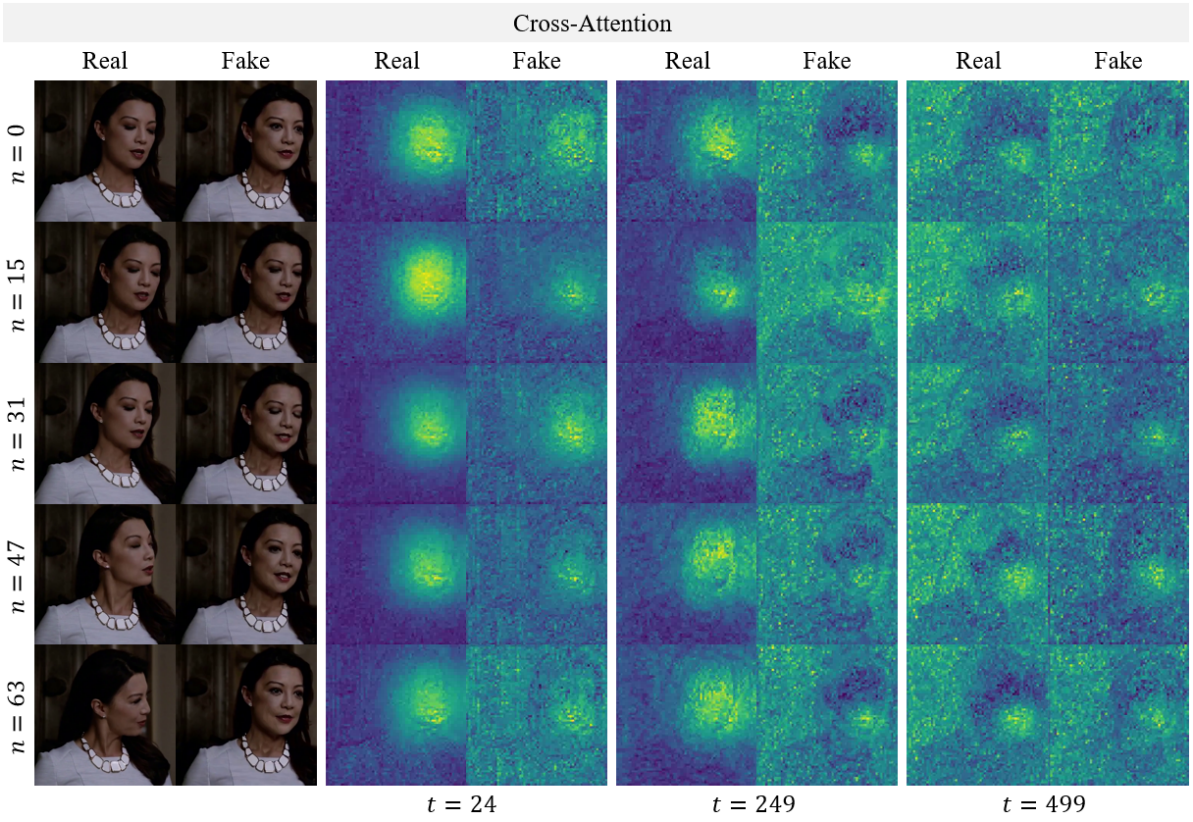
**Analysis on Chosen Timestep.** As reported in Table 6 of the main paper, the performance of X-AVDT consistently improved as the timestep decreased, across cross-attention, temporal-attention, and spatial-attention. This tendency likely arises because features become more informative in earlier diffusion steps (i.e., as $t \to 0$), while features in later steps are more heavily corrupted by noise and thus less discriminative. This observation aligns with prior findings [42, 57, 80, 81, 92, 93], which show that mid-to-early diffusion features provide stronger signals for correspondence, stylization, and segmentation due to reduced noise perturbation and richer structural detail. Therefore we did not conducted experiments on $t > 500$, following the previous research.

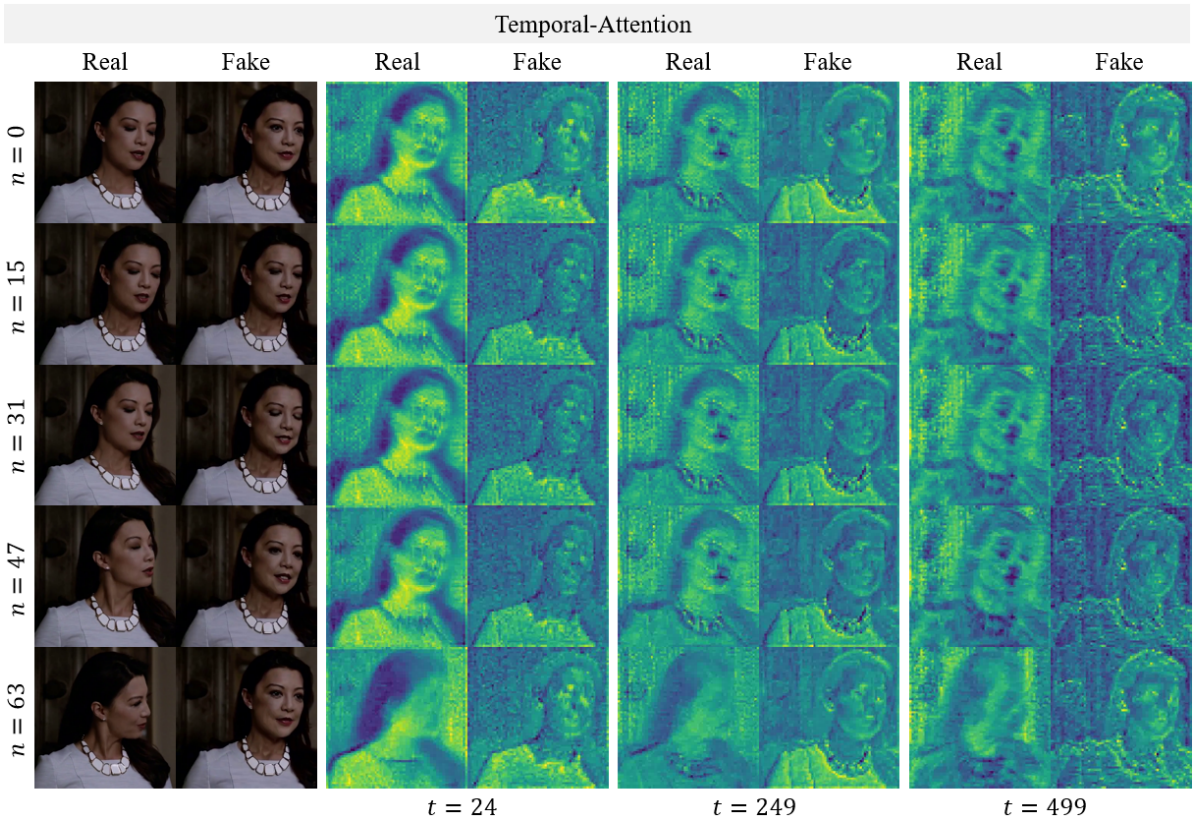
# D. Analysis

This section complements our method by analyzing the overall detection pipeline, and visualizing the internal audio-visual cross-attention maps from the diffusion backbone. We compared our method against a visual-only baseline [11]. For the visualization, we present attention heatmaps for the source and representative generators.

## D.1. Fisher SNR and LDA Margin

We hypothesize that internal audio-visual cross-attention features from large generative models provide a strong discriminative signal for deepfake detection. As shown in Fig-


Cross-Attention
Real
Fake
Real
Fake
Real
Fake
Real
Fake
$n = 0$
$n = 15$
$n = 31$
$n = 47$
$n = 63$
$t = 24$
$t = 249$
$t = 499$


Temporal-Attention
Real
Fake
Real
Fake
Real
Fake
Real
Fake
$n = 0$
$n = 15$
$n = 31$
$n = 47$
$n = 63$
$t = 24$
$t = 249$
$t = 499$

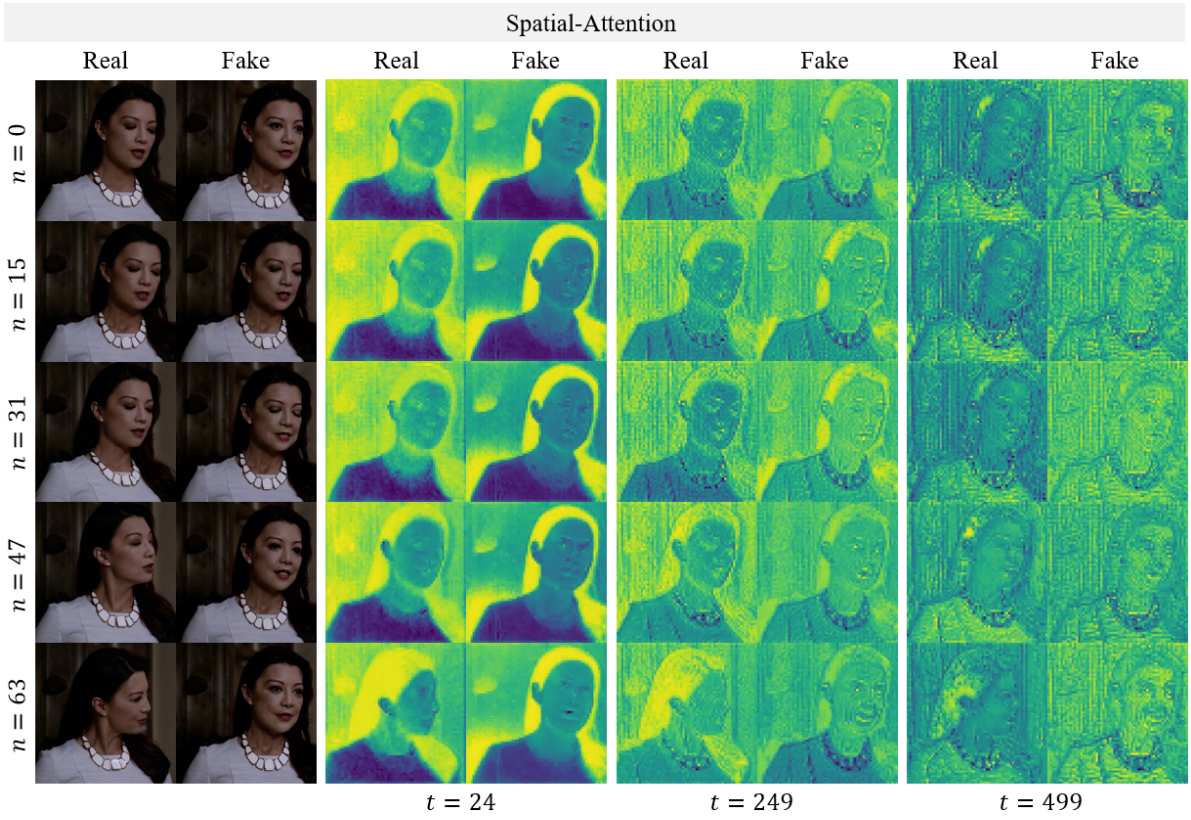



Figure C.3. **Visualization of attention features across diffusion timesteps.**

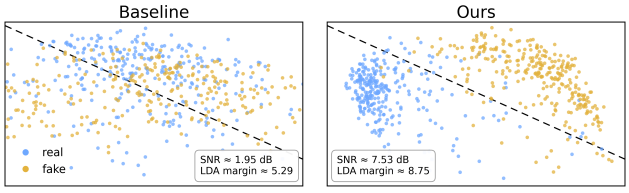



Figure D.1. **PCA embeddings with a shared LDA decision boundary.** Embeddings are extracted from the baseline and from the independently trained X-AVDT, without any fine-tuning.

ure D.1, our method produced noticeably better real/fake separation than a visual only baseline [11]. For a fair comparison, we fit a single linear discriminant analysis (LDA) classifier in the embedding space and project both methods to two dimensions, using the same decision boundary across panels. Measured by Fisher signal-to-noise ratio (SNR) [25] and the LDA margin, performance improves from **1.95dB and 5.29** (baseline) to **7.53dB and 8.75** (ours). This gap suggests that cross-attention captures stable audio-visual correspondence and exposes inconsistencies that generative models fail to reproduce.

## D.2. Cross-Attention Robustness

Figure D.2 quantifies the top-$q$ attention mass coverage within the face ROI: real videos concentrate the top-$q$ mass in a smaller ROI, whereas synthesized videos consistently require coverage of a larger ROI coverage (left). Moreover, the $\Delta$ attention maps reveal a coherent spatial contrast pattern: attention for real videos is concentrated on the mouth and background, while attention for fake videos is more broadly distributed along the face boundary (right). This pattern persists across two different inversion sources, Hallo [90], our backbone generator, and Echomimic [16].

## D.3. Attention Map Visualization

To complement our quantitative results, we visualize internal audio-visual cross-attention maps from the diffusion backbone. As shown in Figure D.3, for each video we extract cross-attention during DDIM inversion, normalize the weights per frame, and average them over time to obtain a single heatmap. We compare the source clip deepfake results from with three representative generators that span different synthesis frameworks: LivePortrait (GAN-based) [30], AniPortrait (diffusion-based) [87], and HunyuanAvatar (flow-matching-based) [15]. Empirically, similar cross-attention patterns are observed across the results from different generator frameworks, indicating that large generative models already provide strong and efficient self-supervised representations well suited for detection. Note that our detector is trained and evaluated on attention features, not on visualized maps, which are provided solely for

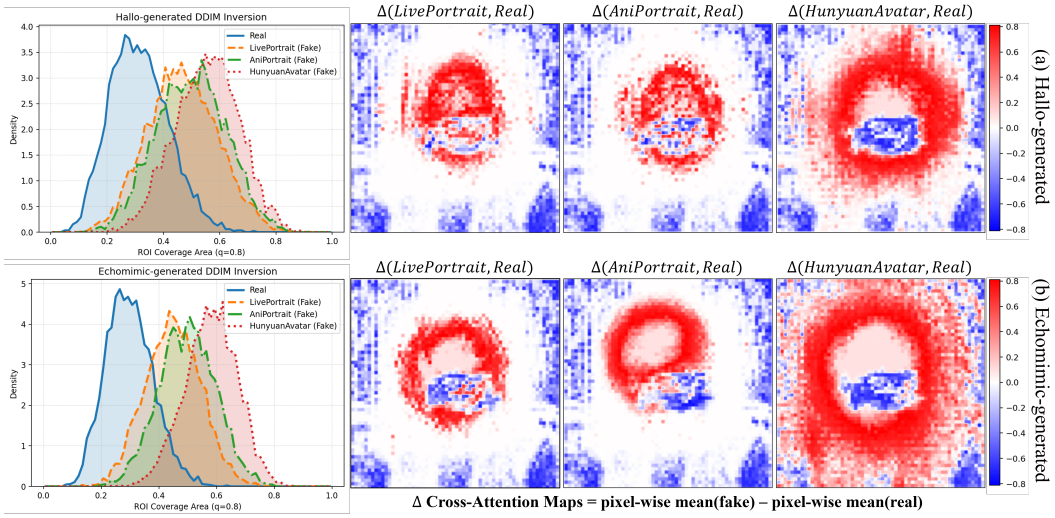



Figure D.2. **Top-$q$ Attention Mass Coverage within the Face ROI (Left) and $\Delta$ Cross-Attention Maps (Right).** In the $\Delta$ maps, red indicates regions with higher fake cross-attention than real ($\Delta > 0$), and blue indicates the opposite ($\Delta < 0$).

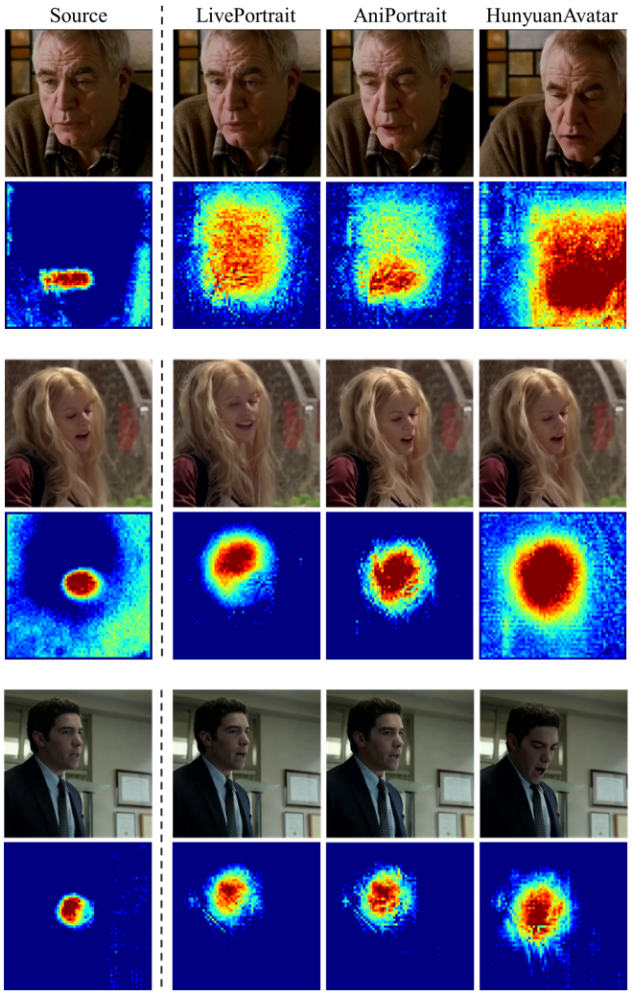



Figure D.3. **Temporally averaged cross-attention heatmaps.**

interpretability. Averaging and normalization for display can introduce information loss, whereas feature vectors are stable.

# E. MMDF Dataset

## E.1. MMDF Construction and Split Protocol

The filtering in MMDF construction corresponds to standard face-detection preprocessing [56] that is widely adopted across facial video datasets and generation pipelines. We only removed clips with inaccurate face tracking to ensure reliable ground-truth pairs, thereby reducing confounding failure cases across all compared detectors rather than favoring X-AVDT. This reduced the candidate set by 6.93% (2,001 clips removed). MMDF is intentionally designed as a strict cross-generator generalization benchmark. Because our goal is to detect forgeries from unseen generation mechanisms, we enforce disjoint generation methods between train and test to avoid overfitting to generator-specific artifacts. We also incorporate a variety of generators, model families, and synthesis methods to maximize the diversity of train-test combinations under this cross-setting.

## E.2. Details of Fake Generators

As mentioned in the main paper, we adopt the Hallo3 dataset released by its authors [20] as the source corpus and employ a curated subset as our real set (see Figure E.1). Then the generators described below synthesize the paired fakes. During preprocessing, all videos are sampled at 25fps to obtain the reference images, and the generated sequences are temporally aligned to their sources for one-to-one pairing, and resized to $512 \times 512$. All fakes are produced by inference only, without any additional training, using the authors' default parameters.

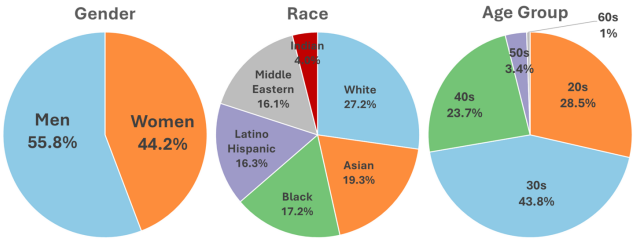



Figure E.1. **Statistics of the MMDF dataset.**

#### E.2.1. Hallo2 [19]

Hallo2 is a diffusion-based, audio-driven portrait image animation model. For the fake samples in the MMDF training set, we use the first video frame as the single reference image and feed the clip's corresponding audio as the driving signal to generate an audio-synchronized talking-head sequence.

#### E.2.2. LivePortrait [30]

LivePortrait is a GAN-based portrait animation method that warps a single reference image according to a driving signal to perform self-reenactment. For the fake samples in the MMDF training set, we use the first frame as the reference image and animate it using the remaining frames as driving images. Note that LivePortrait operates as an image-to-video model without audio driving (i.e., motion is driven solely by image frames).

#### E.2.3. FaceAdapter [33]

FaceAdapter is a face-editing adapter for pretrained diffusion models, targeting face swapping. For the fake samples in the MMDF training set, we randomly select a source identity and a target identity. Leveraging its image-to-image design, we generate swapped frames for the target clip and then pair the synthesized frames with the source audio to produce the final video, thereby preserving the source identity and speech.

#### E.2.4. HunyuanAvatar [15]

HunyuanAvatar is a flow-matching-based audio-driven human animation method. For the fake samples in the MMDF evaluation set, we use the first frame as the source input. We then feed the clip's corresponding audio together with a text prompt generated from the first frame by BLIP-2, OPT-2.7b model, producing an audio-synchronized, text-conditioned talking head sequence.

#### E.2.5. MegActor-$\Sigma$ [91]

MegActor-$\Sigma$ is a diffusion-transformer (DiT)–based portrait animation method with mixed-modal conditioning. For the fake samples in the MMDF evaluation set, we use the first frame as the source input, and feed the source video together with its corresponding audio to generate a self-reenactment sequence.

#### E.2.6. Aniportrait [87]

Aniportrait is a diffusion-based, audio-driven portrait animation method. For the fake samples in the MMDF evaluation set, we use the first video frame as the single reference image and feed the clip's corresponding audio as the driving signal to generate an audio-synchronized talking-head sequence.

### E.3. Input Representation Visualization

Figure D.2.1 and Figure D.2.2 contain samples from our model's input representation, video composite $\psi$ and AV cross-attention feature $\psi$ utilized by our detector. From top to bottom, we show the audio-visual cross-attention feature $\psi$, the original video $x$, the decoded latent DDIM noise map $D(\hat{z}_T)$, the reconstructed video $D(\hat{z}_0)$, and the reconstruction residual $r = |x - D(\hat{z}_0)|$ of the video composite $\phi$. The supplementary video with audio further illustrates temporal dynamics and audio-visual synchronization patterns.

### E.4. MMDF Dataset Visualization

Figures D.3.1–D.3.6 present samples from the curated MMDF dataset used in our experiments. For each identity, the real frames were taken from the source dataset [20] and the fake videos were generated by respective generators. The supplementary video with audio further illustrates temporal dynamics and audio-visual synchronization patterns.

## F. Limitations

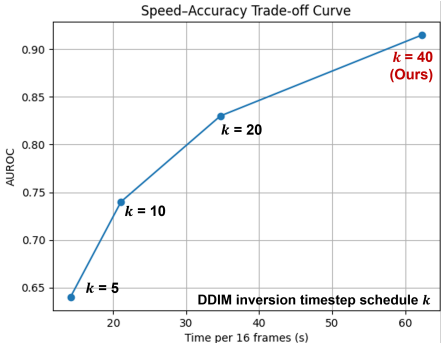



Figure F. **Speed-accuracy trade-off with DDIM inversion steps.**

Figure F indicates a potential limitation of our approach in real-world applications. Performance depends on the number of DDIM inversion timestep schedule $k$ used to extract $\phi$ and $\psi$. While larger $k$ yields more faithful inversion features and improves AUROC, it increases runtime and computational cost. Conversely, smaller $k$ reduces latency but degrades detection accuracy. The incurred cost can be mitigated in future work by adopting fewer step schedules or model distillation.

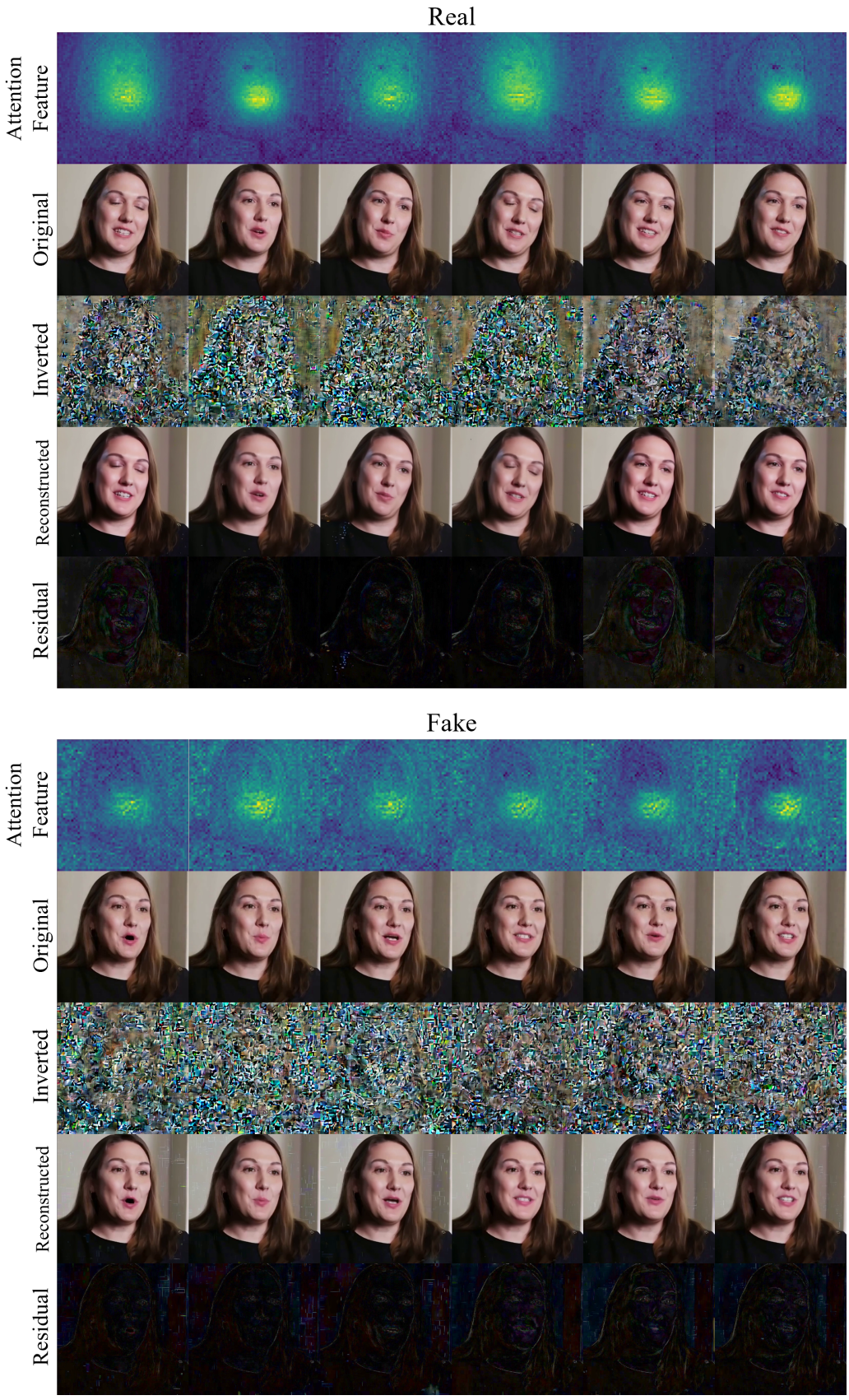



Figure D.2.1. **Qualitative visualization of the input representations $\phi$ and $\psi$.**

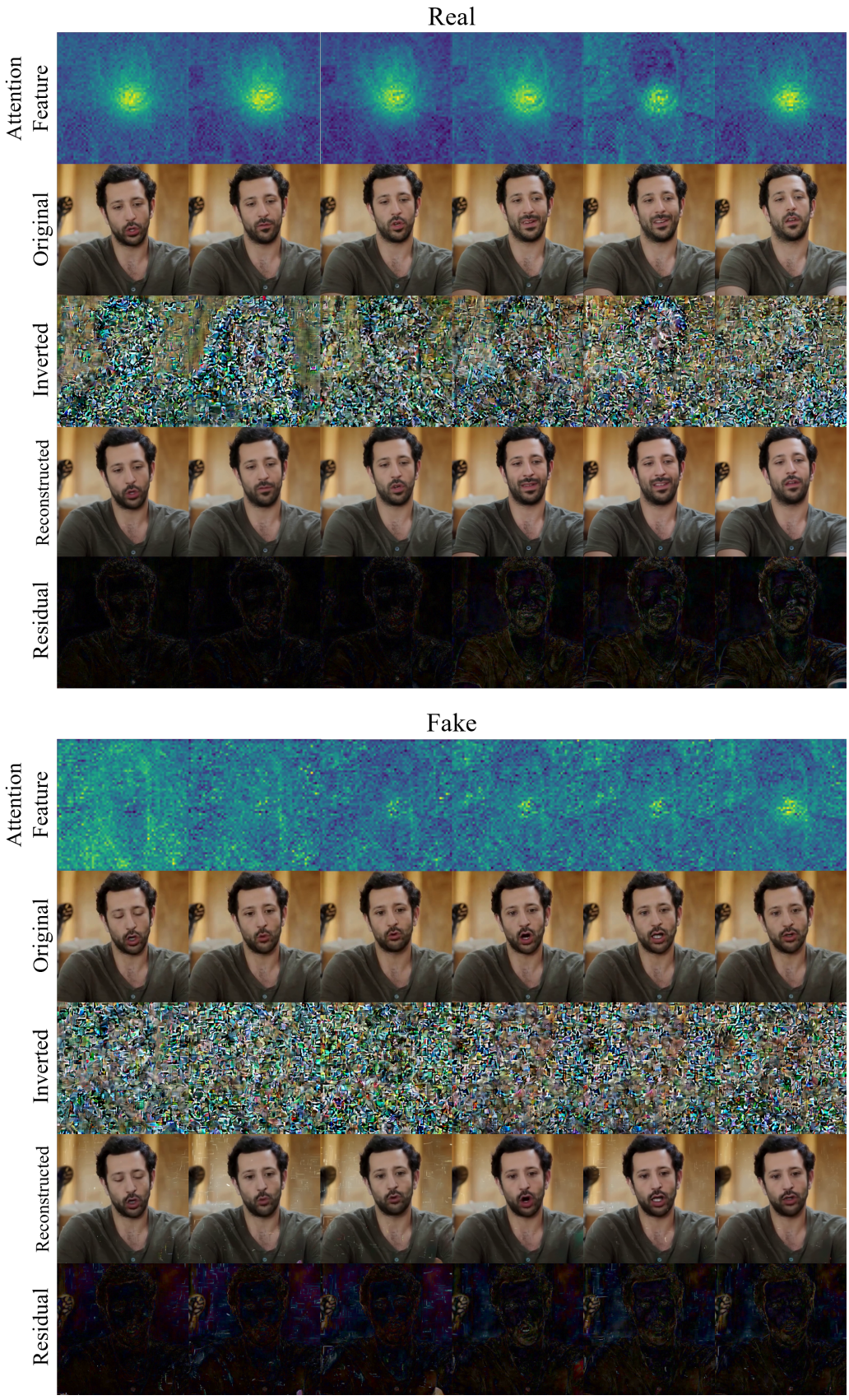



Figure D.2.2. **Qualitative visualization of the input representations $\phi$ and $\psi$.**

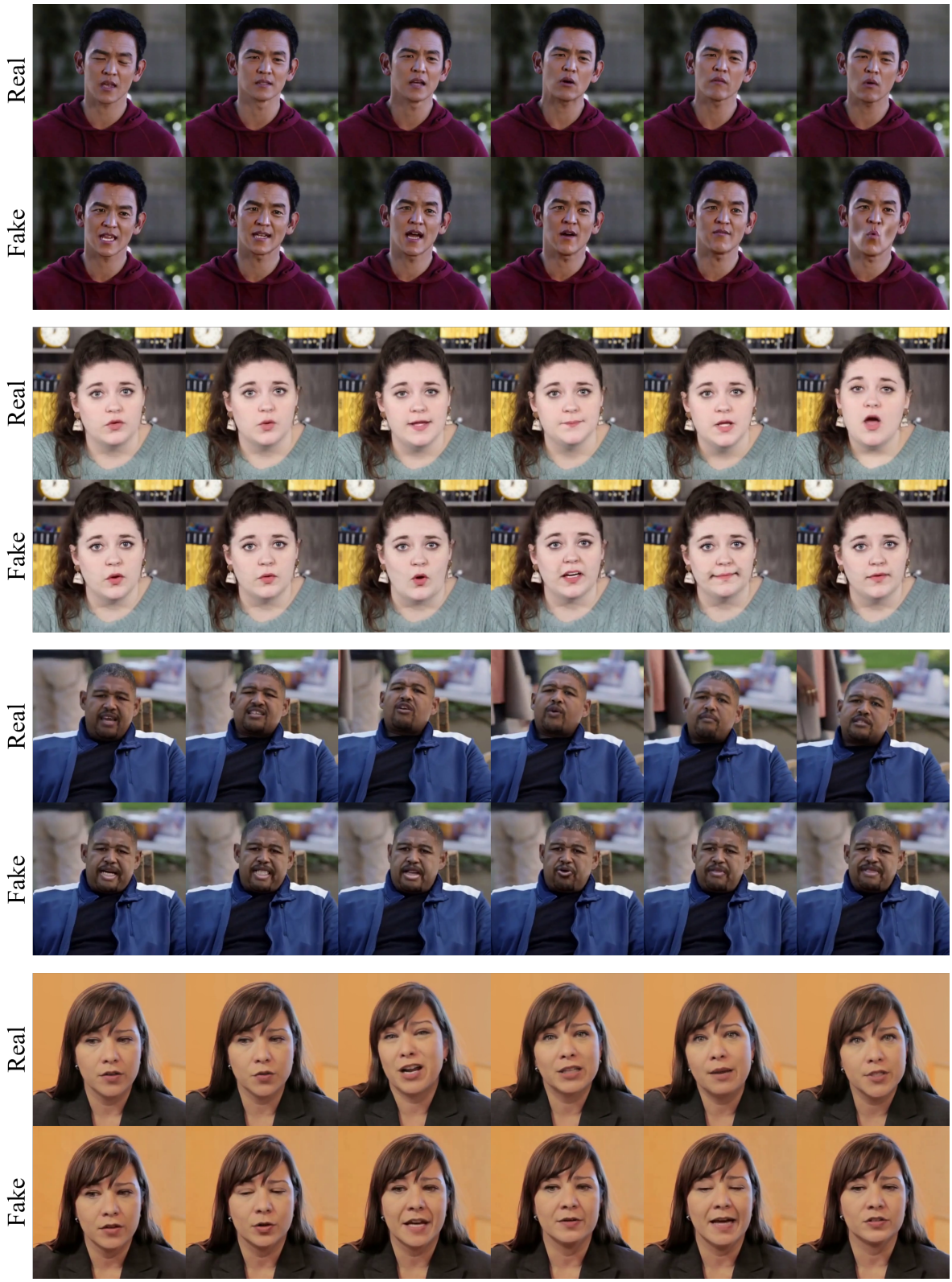



Figure D.3.1. **Qualitative comparison of real and fake videos generated by Hallo2 [19] in the MMDF.**

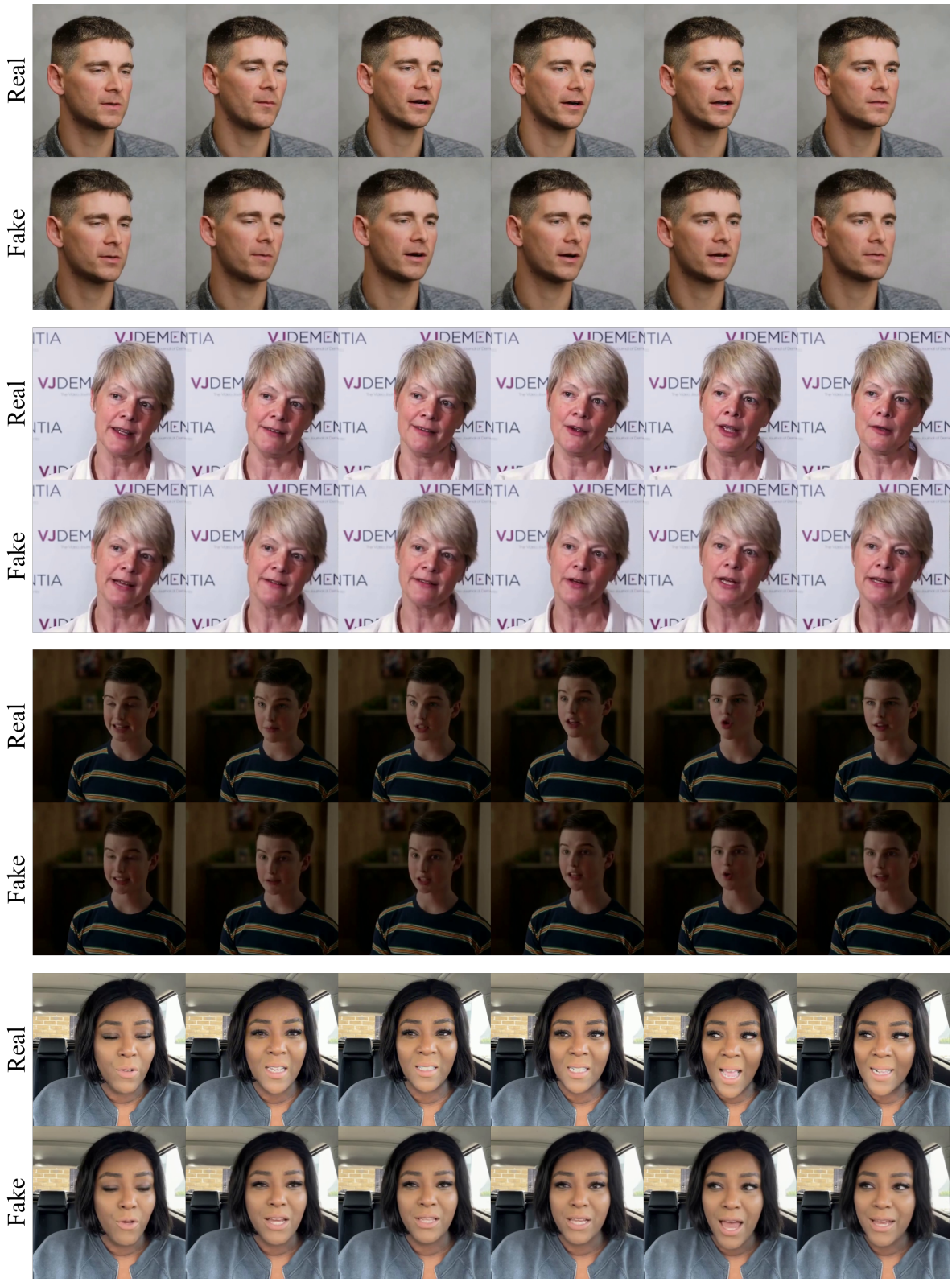



Figure D.3.2. **Qualitative comparison of real and fake videos generated by LivePortrait [30] in the MMDF.**


Figure D.3.3. **Qualitative comparison of real and fake videos generated by FaceAdater [33] in the MMDF.**

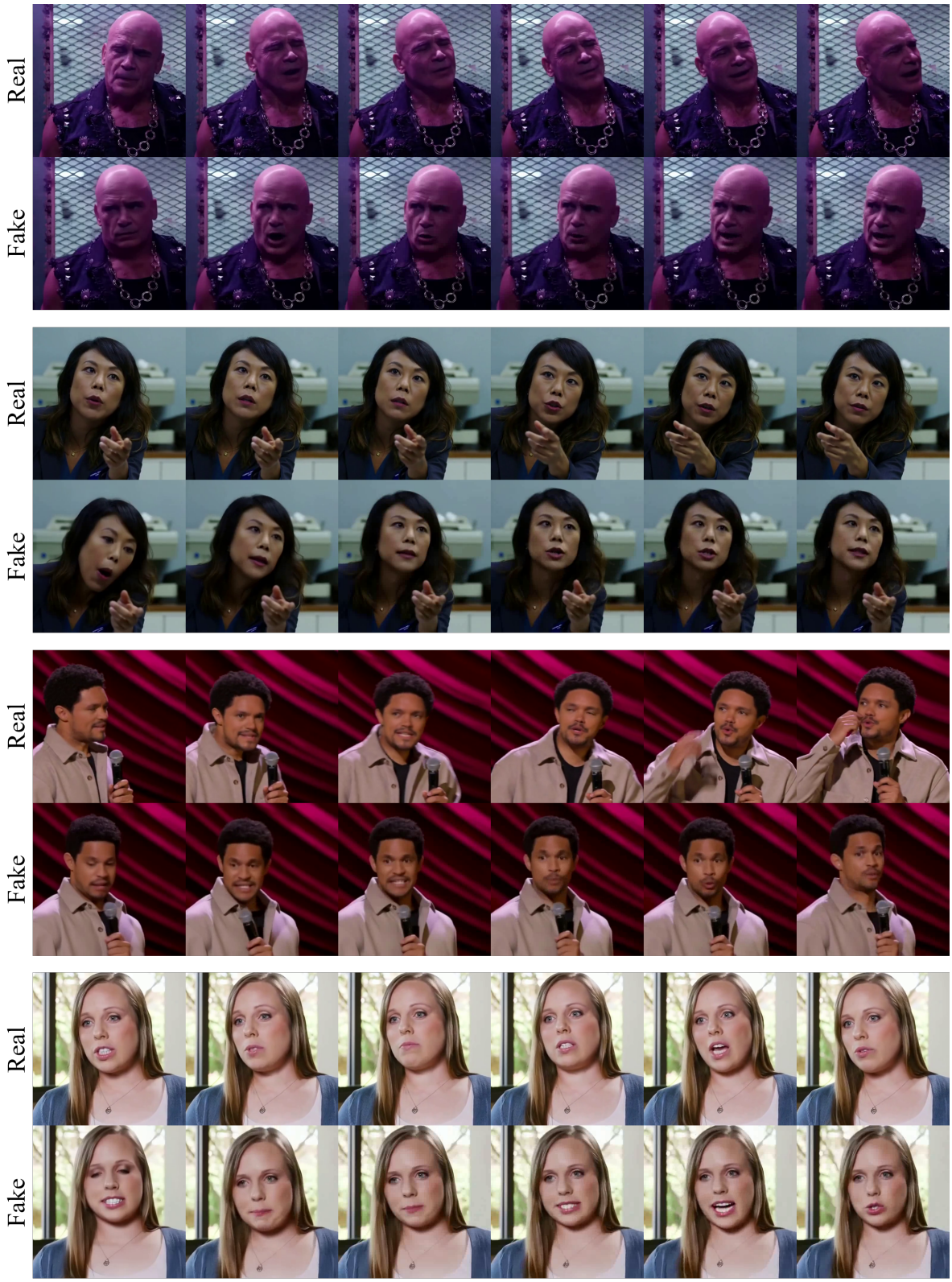



Figure D.3.4. **Qualitative comparison of real and fake videos generated by HunyuanAvatar [15] in the MMDF.**

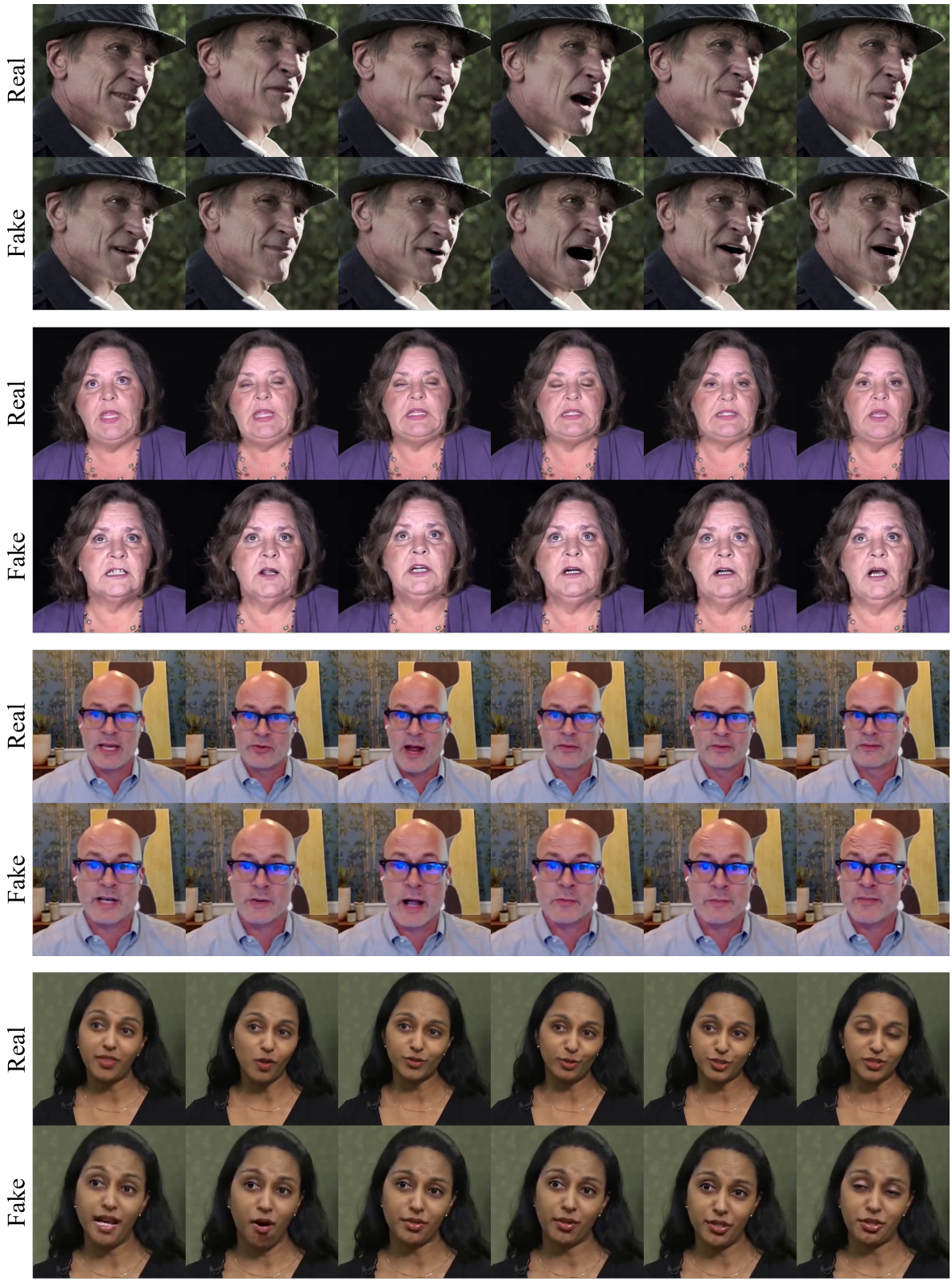



Figure D.3.5. **Qualitative comparison of real and fake videos generated by MegActor-Σ [91] in the MMDF.**

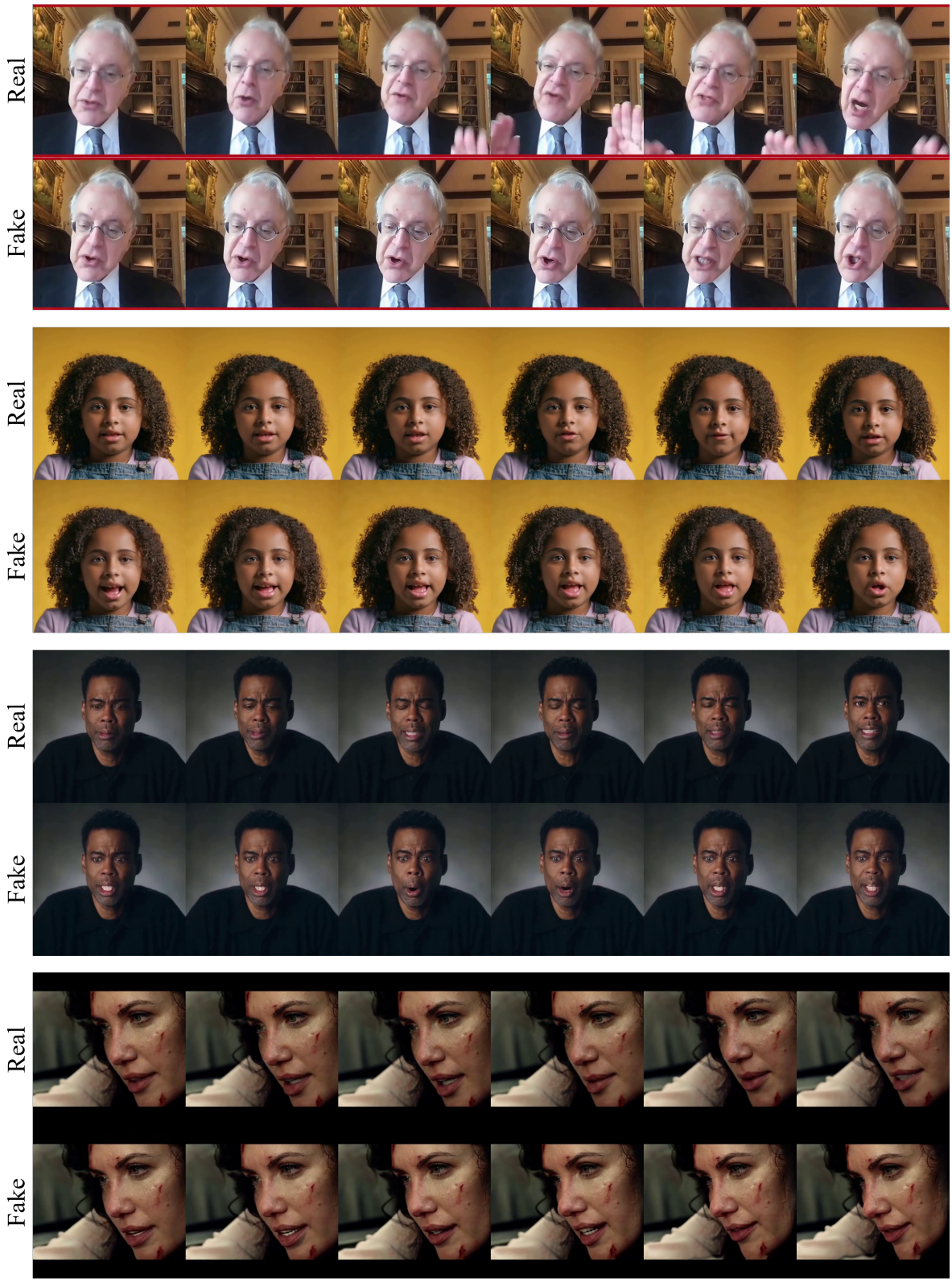



Figure D.3.6. **Qualitative comparison of real and fake videos generated by AniPortrait [87] in the MMDF.**