

# PIA: Deepfake Detection Using Phoneme-Temporal and Identity-Dynamic Analysis

Soumya Kanti Datta\*, Tanvi Ranga\*, Chengzhe Sun, Siwei Lyu  
University at Buffalo, SUNY  
Buffalo, NY, USA

{soumyyak, tanviran, csun22, siweilyu}@buffalo.edu

## Abstract

*The rise of manipulated media has made deepfakes a particularly insidious threat, involving various generative manipulations such as lip-sync modifications, face-swaps, and avatar-driven facial synthesis. Conventional detection methods, which predominantly depend on manually designed phoneme–viseme alignment thresholds, fundamental frame-level consistency checks, or a unimodal detection strategy, inadequately identify modern-day deepfakes generated by advanced generative models such as GANs, diffusion models, and neural rendering techniques. These advanced techniques generate nearly perfect individual frames yet inadvertently create minor temporal discrepancies frequently overlooked by traditional detectors. We present a novel multimodal audio-visual framework, Phoneme-Temporal and Identity-Dynamic Analysis (PIA), incorporating language, dynamic face motion, and facial identification cues to address these limitations. We utilize phoneme sequences, lip geometry data, and advanced facial identity embeddings. This integrated method significantly improves the detection of subtle deepfake alterations by identifying inconsistencies across multiple complementary modalities. Code is available at <https://github.com/skrantidatta/PIA>*

## 1. Introduction

The rapid advancement of generative AI technologies has resulted in an increase in tools capable of creating synthetic media. These tools, in turn, have led to a proliferation of deepfakes, which are AI-created or manipulated videos. The distinctions between authentic and deepfake media become increasingly challenging for human viewers. Although deepfakes involving humans have gained significance in the entertainment sector, they also pose a significant risk to identity, trust, and societal integrity. Recent

events have highlighted the potential identity and security vulnerabilities posed by deepfakes. In February 2024, a multinational corporation incurred a loss of \$25 million due to an employee being deceived by a deepfake impersonation of their chief financial officer and other senior officials. The employee, perceiving it as a legitimate request, transferred funds to a fake account [6]. In another incident, a deepfake impersonator from North Korea deceived KnowBe4, a cybersecurity firm, into employing them in the latter half of 2024 [5]. The ability to deceive a cybersecurity firm demonstrates the remarkable efficacy of these forgeries.

Most existing deepfake detection methods use only one modality, predominantly relying on analysis focused solely on audio or visual signals. There are several works, such as [1], and [29], in which they use multimodal cues but rely on rule-based alignment for audio-visual cues. However, such methods are insufficient in identifying complex manipulations generated by recent developments in generative adversarial networks (GANs) [19] and diffusion-based [15, 24] models, as these models generate high-fidelity facial dynamics and speech-driven articulations that reduce conventional audio-visual alignment discrepancies.

In this work, we develop a new multimodal deepfake detection method, Phoneme-Temporal and Identity-Dynamic Analysis (PIA), to identify audio-visual inconsistency patterns using temporal inconsistencies cues between audio and visual signals. Prior work by Agarwal et al. [1] has demonstrated that such phoneme-viseme mismatches can serve as reliable indicators of manipulation, particularly in the context of deepfakes generated by automated lip-syncing.

We hypothesize that deepfakes insufficiently replicate the sophisticated visemic articulation corresponding to particular phonemes, especially bilabial and rounded vowels such as /m/, /b/, /p/, and /o/. In actual human speech, these phonemes correspond to distinct and reproducible articulatory motions, including complete lip closures for /m/, /b/, and /p/ and a characteristic lip rounding for /o/. Real video sequences exhibit a regular pattern of lip geometry at frames

\*Equal contribution.

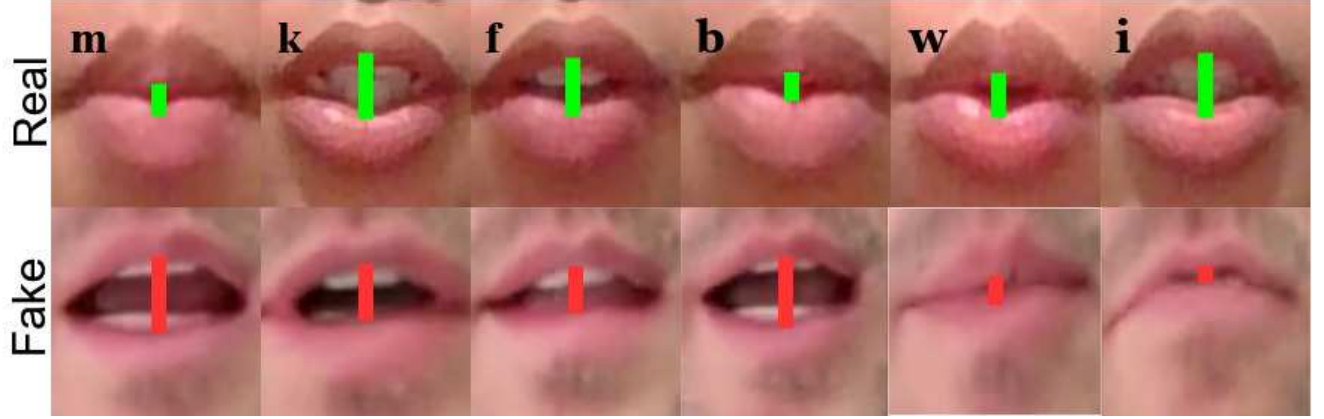


Figure 1. Comparison of lip shapes corresponding to different phonemes. Note that in lip-sync deepfake videos, the degree of lip closure often does not correspond to the sound being pronounced.

temporally synchronized with these phonemes, illustrating the physiological limitations of human articulation. This phenomenon is closely related to the McGurk effect [34], where conflicting auditory and visual speech cues result in a perceptual illusion. It highlights how strongly the human brain relies on audio-visual coherence, especially in speech perception. Additionally, we hypothesize that ArcFace [13] embeddings in real videos exhibit a progressive and consistent temporal progression, marked by smooth transitions between successive frames. In contrast, fake videos, especially those produced by face-swapping methods, frequently exhibit sudden and inconsistent shifts in the embedding space. The  $\ell_2$  distance, or Euclidean distance, is calculated as the square root of the sum of the squared differences between comparable elements of two embedding vectors. This metric measures the frame-to-frame variation in facial identity representation as conveyed by ArcFace. These data indicate that significant abrupt changes in embedding distance can act as a reliable indicator for potential manipulation in an operational deepfake detection system.

Our approach introduces a multimodal architecture directed by a distinct temporal loss function that correlates 14 diverse spoken phoneme letters with lip shape, mouth closure score, and facial identity variation to identify temporal inconsistencies between audio and various facial landmarks. By combining multimodal signals, the model easily discerns the three primary manipulation media: lip-sync, face-swap, and avatar-based deepfakes. We performed comprehensive experiments and ablation studies on two benchmark datasets to assess the effectiveness of our method. The results demonstrate the robustness of our model in identifying in-domain deepfakes.

In summary, our work presents the following primary contributions:

- We highlight the temporal discrepancies between audio-visual cues across 14 distinct phoneme letters, including

bilabials and vowels.

- We present an innovative deepfake detection pipeline that integrates three complementary streams (viseme images, identity embeddings, and lip geometry) through a shared multi-headed attention mechanism and controlled fusion.
- We systematically measure the effect of each modality in controlled ablation studies.
- Our methodology attains an AUC of 98% on the DeepS-pk v2.0 dataset.
- We propose an auxiliary loss that penalises temporal inconsistencies in identity embeddings across successive frames. This loss promotes a more consistent identity representation over time that helps to detect abrupt identity shifts, often observed in face-swap deepfakes.

## 2. Related Work

**Deepfake generation.** As AI technologies advance, more sophisticated and easily accessible deepfake generation tools have become widespread. Deepfake videos fall into two main categories: entire-face synthesis (e.g., face-swaps or talking-head generation, such as avatar deepfakes) and partial manipulation, such as lip-syncing deepfakes, which alter only lip movements to match audio. Early video deepfake generation focused on face-swap techniques such as [7, 28, 37, 44, 49] where the entire face of the original identity is replaced with the target identity. Lip syncing deepfakes overwrite only the mouth region to match new audio. [41] trains a GAN to produce lip movements conditioned on speech features, while [10] uses a three-stage process: expression stabilizer, a lip-sync model, and a face enhancement model to create talking heads. More recent models [30, 36] use diffusion-based models to yield sharper and more temporally coherent lip-syncing deepfakes. Avatar deepfakes animate a single image to produce a full talking-head video to match a target audio or video. [20, 55, 56] use multi-stage control and diffusion modules to extract iden-

tity, motion, voice, and emotion embeddings to produce highly realistic speaking avatars. Such methods have consequently made deepfakes increasingly difficult to detect.

**Deepfake detection.** With deepfake generation methods becoming more sophisticated, detection techniques have advanced in parallel as well. Existing methods often leverage either audio, video, or multimodal signals, each with distinct approaches to feature extraction, mismatch detection, and classification techniques.

Visual-only detectors analyze spatial artifacts and temporal inconsistencies in frame sequences [12, 21, 25, 31, 38, 43, 57]. These models aim to detect inconsistencies in the pixel domain, introduced during the manipulation process. The audio-visual detectors exploit the lip-speech mismatches to detect deepfakes [17, 52, 54]. For example, [1] proposed a multimodal method utilizing phoneme-viseme alignment mismatch to detect deepfake videos. [8] integrates audio, video, and physiological signals, capturing multimodal discrepancies including lip-sync and physiological anomalies like temperature variations and pulse irregularities to detect deepfakes. [40] proposed a two-stage video detector that first self-supervises on real videos to learn intrinsic audio-visual correspondences via contrastive and autoencoding objectives, then fine-tunes on real vs. fake data. Further improving on the previous models, [32] focuses on temporal audio-visual inconsistencies, employing global and local encoders built on Vision Transformer [16] pre-trained on CLIP [42] to identify mismatches. Its classifier utilizes transformer architectures equipped with a dynamic attention module to detect deepfakes.

Our proposed pipeline advances beyond these existing methods by integrating a richer set of multimodal inputs, including phonemes from WhisperX [2] combined with phonemizer, visemes and lip landmarks from MediaPipe [33], and facial identity embeddings from ArcFace [13]. We introduce nuanced mismatch signals such as soft lip-closure scores, viseme-phoneme mismatches, and ArcFace drift to capture subtle articulatory and identity-based discrepancies. This comprehensive approach demonstrates improved performance in intra-dataset performance and cross-manipulation generalization.

### 3. Backgrounds

Our method is based on observations of the temporal inconsistencies existing in various types of deepfakes, and these observations provide the foundation for the model architecture and detection strategies discussed in subsequent sections.

**Phonemes Visemes Mismatch Pattern.** Lip-sync deepfake videos may have temporal inconsistencies due to the mismatch of phonemes and visemes [1]. To demonstrate this phenomenon, we selected a curated set of 14 phonemes based on their distinct articulatory features, high visual

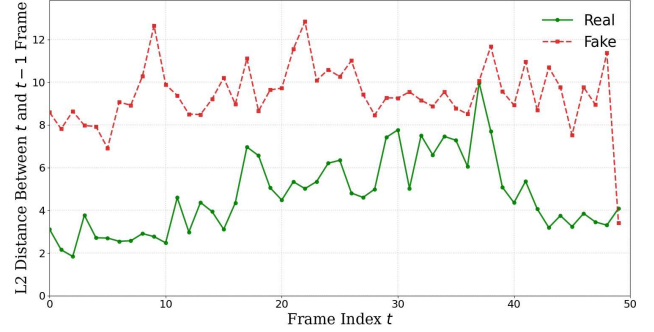


Figure 2. Temporal drift in ArcFace embeddings between consecutive frames for a real video (green) and its face-swap deepfake counterpart (red). Higher drift in initial frames and consistently higher overall drift in fake samples indicates temporal identity inconsistencies.

salience, and coverage of key phoneme classes important for audiovisual speech modeling and deepfake detection. This subset spans a diverse range of phonetic categories:

- **Bilabials:** /p/, /b/, /m/ — require full lip closure and are visually distinguishable.
- **Labiodentals:** /f/, /v/ — involve lip-to-teeth contact, producing clear articulatory movements.
- **Alveolars:** /t/, /s/ — frequent in natural speech and involve rapid tongue and jaw movement.
- **Velars:** /k/ — exhibit distinct mouth shapes with less lip movement but unique viseme cues.
- **Approximants:** /w/, /r/ — require lip rounding or protrusion, easily detected in lip shapes.
- **Vowels:** /i/, /æ/, /o/ — include high front, low front, and mid back vowels, capturing open-mouth gestures.
- **Postalveolar:** /ʃ/ — involves noticeable lip rounding and is visually distinct.

These phonemes are chosen to maximize the ability to detect mismatches between visual articulation and audio content, particularly in manipulated videos where viseme-phoneme alignment is disrupted. Phonemes with low visual distinguishability (e.g., glottals or unstressed vowels) and silences were excluded to avoid noisy or uninformative supervision. We observe that lip-sync manipulations often fail to maintain alignment between speech articulation and lip movements. This discrepancy is mostly observed in labial, labiodental, and vowel phonemes such as /p/, /b/, /m/, /f/, /v/, and /o/, which require precise lip closure or openness.

As presented in Fig. 1, in several lip-sync generated videos, the lip geometry failed to match expected phoneme articulations, particularly during high closure phonemes. These mismatch signals were consistently detected using MediaPipe-based lip geometry analysis.

**Temporal Drift in Identity Embeddings.** We observe that for real videos, ArcFace-based identity embeddings transi-

tion smoothly across the frames. In contrast, fake videos, especially generated by face-swapping manipulation techniques, display sharp, irregular embedding shifts, suggesting that identity preservation across time is a useful indicator of face-swaps.

As seen in Fig. 2, a plot of  $\ell_2$  distances between consecutive ArcFace embeddings reveals that real videos maintain low and stable distances between 2-6  $\ell_2$  distance while fake videos exhibit sharp spikes at 8-12  $\ell_2$  distance in the beginning. This supports the use of identity drift as a reliable indicator for early-stage manipulation detection.

## 4. Method

Our proposed method integrates a multimodal feature extraction pipeline with a unified deepfake detection model that simultaneously analyzes phoneme articulation, lip geometry, viseme appearance, and identity cues. The model employs a 3D convolutional network [47] with a pre-trained EfficientNet-B0 [46] backbone for extracting visual information from mouth-region images, along with a multihead attention [48] mechanism to efficiently capture temporal and modality-specific relationships. Phonemes obtained by WhisperX [2] and aligned with wav2vec2 [51], are used as an active pre-processing filter to select temporally meaningful frames when phoneme articulation is visually significant. This architecture is designed to detect barely noticeable discrepancies caused by lip-sync and face-swap manipulations by studying cross-modal mismatches between audio and visual streams.

### 4.1. Feature Extraction

For each input video, we perform a structured pre-processing procedure to extract synchronized audio-visual and geometric features required for downstream deepfake detection. This process includes four key stages: audio extraction and phoneme alignment, visual feature extraction, facial identity embedding, and frame-level alignment.

**Audio Extraction and Phonemes Alignment.** We begin the process by extracting the raw waveform from every video, using FFmpeg [18] and resampling the audio to a 16 kHz mono-channel format in order to preserve uniformity among samples. Speech transcription uses the WhisperX [2] large-v2 model, producing word-level segments accompanied by accurate timestamps. The segments are subsequently transformed into sequences of International Phonetic Alphabet (IPA) phonemes via the phonemizer [4] package. We use a wav2vec2-based alignment model [51] to accurately match phonemes with the audio waveform, improving the phoneme timestamps at sub-word resolution. Ultimately, we associate each video frame with a corresponding phoneme label by interpolating its timestamp inside the matched phoneme intervals, therefore assuring synchronised frame-level phoneme annotation.

**Visemes Feature Extraction.** Each frame is analysed using MediaPipe [33] FaceMesh, which identifies 468 facial landmarks with exceptional spatial precision. To concentrate on articulatory dynamics pertinent to speech, we extract a subset of 27 lip-related landmarks that correspond to critical places along the outer and inner contours of the mouth. We derive geometric descriptors from these locations to quantify lip movement and morphology. Lip height is the vertical distance between the central landmarks of both the upper and lower lips, whereas lip width is the horizontal distance between the left and right corners of the lips. Utilising these two measurements, we determine the mouth aspect ratio (MAR), defined as the ratio of lip height to lip width.

$$\text{Aspect Ratio} = \frac{\text{lip\_height}}{\text{lip\_width} + \varepsilon} \quad (1)$$

where  $\varepsilon$  is a small constant (e.g.,  $1 \times 10^{-6}$ ) added to prevent division by zero. This ratio functions as a crucial measure of mouth openness, with elevated values associated with vowel-like articulations and diminished values generally noted during lip closure or consonant production. These geometric indicators are especially valuable for discerning phoneme-specific articulation patterns and for recognizing discrepancies in deepfake videos. Bilabial phonemes such as /m/, /b/, and /p/ are anticipated to demonstrate low mouth aspect ratio (MAR) values owing to complete lip closure, but open vowels like /a/ or /i/ yield greater mouth aspect ratio (MAR) values due to vertical mouth extension.

**Facial Identity Embedding.** We use the ArcFace [13] model from InsightFace [14] to obtain frame-level speaker identity representations. ArcFace is a state-of-the-art face recognition model that projects facial features into a 512-dimensional hypersphere space using an additive angular margin loss, which ensures high inter-class separability and intra-class compactness. Each frame of the input video is independently processed by the ArcFace model to generate a 512-dimensional identity embedding. These embeddings encapsulate advanced facial characteristics, including skeletal structure, facial ratios, and expression-invariant traits, making them well suited to detect the slight yet discernible identity drift between frames caused by face-swap induced temporal anomalies. By examining the coherence of ArcFace embeddings across successive frames, the model can identify identity drift that may not be evident at the pixel level. These embeddings are retained for two key purposes: (1) as one of the input modalities in the multimodal fusion model, where they are encoded and integrated with viseme and geometric features, and (2) for auxiliary supervision via a temporal consistency loss, which penalizes abrupt changes in identity features across adjacent frames and encourages the model to learn stable identity dynamics typical of real videos.

**Multi modal Representation.** To train the model using



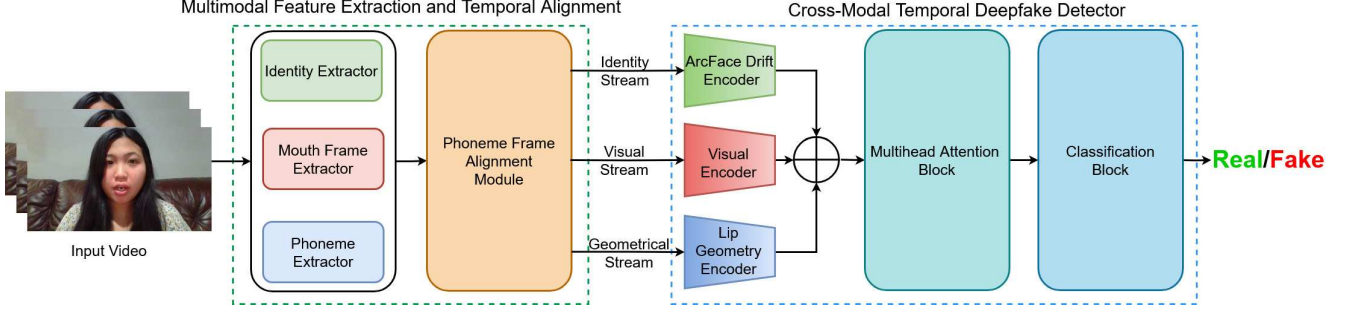


Figure 3. End-to-end pipeline of our proposed PIA model. It consists of (1) Multimodal Feature Extraction and Temporal Alignment, and (2) Cross-Modal Temporal Deepfake Detector.

frame-level multi-modal cues, we construct a dataset in which each phoneme label is temporally aligned with its corresponding audio segment and drawn from a predefined set of 14 phonemes. For each phoneme in this set, five frames are uniformly sampled from the video instances where that phoneme appears. For each of these frames, we extract three input streams: (1) viseme image crops, (2) identity embeddings generated using ArcFace, and (3) geometric descriptors computed from lip landmarks. Non-linguistic tokens, including silences, noise, and pauses, are omitted. Phoneme tokens are used to filter frames according to a carefully selected vocabulary of 14 visually and articulately distinct elements. This selection guarantees that only frames associated with prominent speech articulations are preserved for model training, hence improving the quality of visual, geometric, and identification aspects.

## 4.2. Model Architecture

Our multistream deepfake detection model integrates three modalities: (1) lip geometry descriptors, (2) viseme image crops, and (3) ArcFace identity embeddings. Each stream is encoded independently using a dedicated encoder, and the resulting features are fused via multi-head attention based pooling [48] for final classification, as shown in Fig. 3. The ArcFace drift encoder is a multilayer perceptron that encodes the 512-dimensional ArcFace identity embeddings for each phoneme group. The visual encoder comprises a 3D convolutional network designed to capture spatio-temporal inconsistencies across the five-frame sequence associated with each phoneme. The resulting outputs are temporally averaged and subsequently processed by a pretrained EfficientNet-B0 backbone to extract higher-level visual representations. The lip geometry encoder is a multilayer perceptron that encodes the mouth aspect ratios computed from lip landmarks across frames.

To integrate multimodal features and summarize temporal dynamics, we concatenate modality-specific embeddings (geometry, visual, identity) into a unified feature:

$$\mathbf{f}_t \in \mathbb{R}^{3d}, \quad \mathbf{f}_t = \mathbf{g}_t \oplus \mathbf{v}_t \oplus \mathbf{a}_t \quad (2)$$

The resulting sequence  $\{\mathbf{f}_t\}_{t=1}^T$  is then summarized using multi-head attention pooling to obtain a global video-level representation. Each learnable query attends over the temporal sequence to produce head-specific summaries, which are averaged to yield a compact representation:

$$\mathbf{z} = \frac{1}{H} \sum_{h=1}^H \sum_{t=1}^T \alpha_{h,t} \cdot \mathbf{f}_t' \quad (3)$$

where  $\alpha_{h,t}$  is the attention score assigned to each input frame  $t$  by attention head  $h$ .

## 4.3. ArcFace Temporal Consistency Loss

To enforce temporal coherence in the identity representation, we introduce an *ArcFace Temporal Consistency Loss*, which penalizes abrupt or implausible changes in the ArcFace facial identities across consecutive frames.

Let  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_T \in \mathbb{R}^d$  denote the ArcFace embeddings for a video sequence of  $T$  frames. For each adjacent pair of time steps  $(t, t+1)$ , we compute the cosine similarity:

$$s_t = \cos(\mathbf{a}_t, \mathbf{a}_{t+1}) = \frac{\mathbf{a}_t \cdot \mathbf{a}_{t+1}}{|\mathbf{a}_t| \cdot |\mathbf{a}_{t+1}|} \quad (4)$$

We define the identity deviation as  $1 - s_t$ , which reflects the degree of identity shift between frames. To ignore irrelevant frames (e.g., during silence), we apply a binary mask  $m_t \in \{0, 1\}$ , where  $m_t = 1$  indicates a non-silent frame.

The overall ArcFace temporal consistency loss is then defined as:

$$\mathcal{L}_{\text{arcface}} = \frac{\sum_{t=1}^{T-1} (1 - \cos(\mathbf{a}_t, \mathbf{a}_{t+1})) \cdot m_t \cdot m_{t+1}}{\sum_{t=1}^{T-1} m_t \cdot m_{t+1} + \epsilon} \quad (5)$$

where  $\epsilon$  is a small constant added for numerical stability. This loss encourages smooth identity transitions in real videos and helps expose temporal inconsistencies introduced by manipulations in deepfakes. The final loss is calculated as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{arcface}} \quad (6)$$

where  $\mathcal{L}_{\text{CE}}$  is the cross-entropy loss, and  $\lambda$  is a weighting coefficient on the ArcFace temporal consistency loss.



Figure 4. Example frames from the FakeAVCeleb dataset (left) and the DeepSpeak v2.0 dataset (right). The top row shows frames from lip-sync deepfakes, while the bottom row presents frames from face-swap deepfakes. Notable differences in native resolution and visual quality are apparent. For comparison, face regions from FakeAVCeleb were enlarged to match the scale of DeepSpeak v2.0 samples.

## 5. Experiments

### 5.1. Experimental settings

**Datasets:** In this paper, the experiments are performed on two datasets, namely FakeAVCeleb [27] and DeepSpeak v2.0 [3]. The FakeAVCeleb dataset [27] consists of 20,000 samples, out of which 19,500 are deepfake videos and 500 are real videos of resolution  $224 \times 224$ . We partitioned the dataset into five categories similar to [39]:

- FVRA-WL: FakeVideo-RealAudio-Wav2Lip [41]
- FVFA-FS: FakeVideo-FakeAudio-FaceSwap [28]
- FVFA-GAN: FakeVideo-FakeAudio-FaceSwapGAN [37]
- FVFA-WL: FakeVideo-FakeAudio-Wav2Lip [41]
- RVFA: RealVideo-FakeAudio

Following [39], we used 70% of the dataset to train and validate our model, and the remaining 30% to test our model. Since our model is designed exclusively for fake-video detection, we exclude the RealVideo-FakeAudio category from both the training and test sets.

The DeepSpeak v2.0 dataset [3] consists of 9,376 real videos and 7,209 deepfake videos, such as face-swapping, lip-syncing, and avatar-based fake videos. The real videos in this dataset are of two different resolutions,  $640 \times 480$  and  $1280 \times 720$ . The fake videos are found with three different resolutions:  $640 \times 480$ ,  $512 \times 512$ , and  $1280 \times 720$ . As provided by the dataset, the videos are split into training and testing subsets in an 80:20 ratio, respectively. We further

divide the test set into three categories based on the type of deepfakes as Face-swap, Lip-sync, and Avatar.

**Implementation Details:** We use 14 distinct phonemes in the proposed architecture and construct the dataset by aligning phonetic alphabets extracted from WhisperX [2] with viseme crops and metadata on a per-frame basis. Mouth crops are resized to  $112 \times 112$  pixels and normalised. The model is trained utilising cross-entropy loss with label smoothing and an auxiliary ArcFace temporal consistency loss. For all experiments we use the Adam optimizer with a learning rate of  $3 \times 10^{-4}$ , weight decay of  $1 \times 10^{-5}$ , and an ArcFace temporal consistency loss coefficient  $\lambda = 0.1$ . For our multi-head attention module, we use 4 heads. We train our model for 25 epochs with a batch size of 16 using PyTorch 2.6.0 with CUDA 12.4.

**Evaluation Metrics:** The model is evaluated using three widely used metrics, including Average Precision (AP), Area Under the Receiver Operating Characteristic Curve (AUC) scores, and Accuracy (ACC) scores. We denote percentage points as %-pts.

### 5.2. Results

**Performance on FakeAVCeleb Dataset.** In line with prior work [39], we train our model on the FakeAVCeleb [27] training split and evaluate it on the FakeAVCeleb test split. As shown in Table 1, our proposed method attains the highest results, achieving ACC and AUC scores of 98.7% and 99.8%, respectively, thereby surpassing all baseline models. For the PIA\_RVFA evaluation, the RVFA category is added back in the test set only and excluded from training. Although our model is designed solely for fake-video detection, it achieves 98% accuracy and a 98.2% AUC inclusive of this category.

In Table 2, similar to [39], we assess the model’s ability to generalize to videos manipulated by a method that was not included in the training set. We use four categories: FVRA-WL, FVFA-FS, FVFA-GAN, and FVFA-WL for this experiment. For each of the four categories, we partition the dataset into 70% training and 30% testing with no overlap. During training, we withhold one category entirely and use it solely for testing, thus rotating through all categories in turn. It can be observed that our proposed approach is able to achieve the highest performance compared to the state-of-the-art model for all categories. For FVRA-WL, compared to LipForensics [21], the best-performing baseline model in terms of AP, which achieved 97.8% AP, our approach increases the performance by 2.1%-pts, and when compared to AVFF [39], the top model in terms of AUC with a 98.2% AUC, we improve AUC by 0.9%-pts. In terms of overall average performance (AVG-FV), our method outperforms the best baseline model AVFF [39] by 1.4%-pts in AP and 0.3%-pts in AUC.

Method	Modality	ACC(%)	AUC(%)
Xception [43]	V	67.9	70.5
LipForensics [21]	V	80.1	82.4
FTCN [57]	V	64.9	84.0
CViT [50]	V	69.7	71.8
RealForensics [22]	V	89.9	94.6
Emotions Don't Lie [35]	AV	78.1	79.8
MDS [11]	AV	82.8	86.5
AVFakeNet [26]	AV	78.4	83.4
VFD [9]	AV	81.5	86.1
AVoID-DF [52]	AV	83.7	89.2
AVFF [39]	AV	98.6	99.1
PIA_RVFA (Ours)	AV	98.0	98.2
PIA (Ours)	AV	<b>98.7</b>	<b>99.8</b>

Table 1. **Performance on FakeAVCeleb Test Dataset.** We benchmark our method against baseline models on the FakeAVCeleb dataset using a 70%/30% train-test split. The best results are highlighted in bold.

Our model’s superior performance compared to state-of-the-art methods can be attributed to its capability to concurrently represent phoneme articulation based visual appearance, geometric consistency, and identification cues through a unified multimodal architecture. The use of cross-modal fusion, attention-based temporal pooling, and auxiliary ArcFace temporal consistency loss enables the model to capture subtle spatiotemporal discrepancies that are often ignored by unimodal or weakly fused baselines.

**Performance on Deepspeak v2.0 Dataset.** Here we train and evaluate our model on the Deepspeak v2.0 dataset [3], which offers higher visual quality than FakeAVCeleb as shown in Fig. 4. We use the provided train/test split to evaluate our model. To the best of our knowledge, we are the first to benchmark on this Deepspeak v2.0; prior work [53] has only evaluated Deepspeak v1.0, which lacks avatar-based deepfakes. Their method [53] got an AUC score of 92.01% on Deepspeak v1.0. Our proposed approach was able to achieve an AUC score of 98.06% on Deepspeak v2.0. This shows that our model is capable of detecting high-quality deepfakes as well. The results are shown in Table 3.

### 5.3. Ablation Analysis

To evaluate the contribution of each component within our proposed framework, we present our ablation analysis in Table 3. We use the Deepspeak v2.0 dataset [3] for ablation analysis, since the Deepspeak v2.0 dataset has a higher visual quality as compared to the FakeAvCeleb dataset [27] as shown in Fig. 4. All the models are trained on the training set of the Deepspeak v2.0 dataset. We report AUC scores for each ablation across the Lip-sync, Face-swap,

and Avatar test subsets, as well as for the combined test set referred to as Global in the Table 3. We denote our full model as PIA. Here, “w/o\_vi” refers to excluding viseme image embeddings, “w/o\_geom” refers to excluding the lip geometry stream, “w/o\_EB0” refers to excluding the EfficientNet-B0 [46] CNN backbone, and “w\_ph” refers to including the one-hot encoded phonemes data stream.

**Excluding Visemes (PIA\_w\_ph\_w/o\_vi)** We train the model with the one-hot encoded phonemes and remove the visemes images input feature in order to assess the impact of visual cues on detection performance. It can be observed that the AUC falls by 30.8%-pts, 34.35%-pts, 33.03%-pts, and 32.49%-pts for Lip-sync, Face-swap, Avatar, and Global test subsets, respectively. These results highlight the critical role of visual image cues from the lip region in capturing the inconsistencies.

**Excluding Lips Geometry (PIA\_w\_ph\_w/o\_geom)** In this experiment, we train the model with the one-hot encoded phonemes and remove the lip geometry stream to evaluate the contribution of geometric features in deepfake detection. By removing this component, we assess the model’s reliance on temporal lip shape variations for capturing subtle inconsistencies in lip-sync manipulations. The results show a slight drop in AUC of 0.93%-pts, 4.91%-pts, 0.99%-pts, and 1.57%-pts for Lip-sync, Face-swap, Avatar, and Global test subsets, respectively. This suggests that lips geometry provides complementary information.

**Excluding ArcFace Embeddings (PIA\_w\_ph\_w/o\_arc)** In this experiment, we train the model with the one-hot encoded phonemes and exclude the ArcFace identity embeddings. By eliminating this stream, we assess the model’s capacity to identify temporal inconsistencies without identity-based signals. Here, the AUC dropped by 0.6%-pts, 0.48%-pts, 1.08%-pts, and 1.04%-pts for Lip-sync, Face-swap, Avatar, and Global test subsets, respectively.

**Including One-Hot Encoded Phonemes (PIA\_w\_ph)** In this ablation, we train the model by introducing the one-hot encoded phonemes along with cross-modal attention applied to ArcFace facial embeddings, without removing any other module. These one-hot encoded phoneme features are fused with visual appearance and lip geometry streams to form the phoneme-infused baseline. Here the AUC dropped by 0.29%-pts, 3.93%-pts, 1.33%-pts, and 1.40%-pts for Lip-sync, Face-swap, Avatar, and Global test subsets, respectively. These results suggest that while phonemes are useful for selecting relevant frames with aligned viseme images, lip geometry, and identity features, their inclusion as a fused input stream may introduce noise, thereby limiting the model’s discriminative capacity.

**Excluding EfficientNet-B0 (PIA\_w/o\_EB0)** Here we train the model with frozen pretrained RESNET-18 [23] embeddings in place of the EfficientNet-B0 [46] backbone module to assess the impact of fixed visual representations on model

Method	Modality	FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		AVG-FV	
		AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)
Xception [43]	V	88.2	88.3	92.3	93.5	67.6	68.5	91.0	91.0	84.8	85.3
LipForensics [21]	V	97.8	97.7	99.9	99.9	61.5	68.1	98.6	98.7	89.4	91.1
FTCN [57]	V	96.2	97.4	<b>100.0</b>	<b>100.0</b>	77.4	78.3	95.6	96.5	92.3	93.1
RealForensics [22]	V	88.8	93.0	99.3	99.1	99.8	99.8	93.4	96.7	95.3	97.1
AV-DFD [58]	AV	97.0	97.4	99.6	99.7	58.4	55.4	<b>100.0</b>	<b>100.0</b>	88.8	88.1
AVAD (LRS2) [17]	AV	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
AVAD (LRS3) [17]	AV	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	91.3	92.8
AVFF [39]	AV	94.8	98.2	<b>100.0</b>	<b>100.0</b>	99.9	<b>100.0</b>	99.4	99.8	98.5	99.5
PIA (Ours)	AV	<b>99.9</b>	<b>99.1</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>99.9</b>	<b>99.8</b>

Table 2. **Cross-manipulation evaluation.** We evaluate the model’s performance by leaving out one category for testing while training on the rest, using a 70%/30% train–test split across four manipulation types (FVRA–WL, FVFA–FS, FVFA–GAN, FVFA–WL) on the FakeAVCeleb dataset. Each column heading indicates the held-out test category. The best results are highlighted in bold.

Dataset	Lip-sync	Face-swap	Avatar	Global
PIA_w_ph_w/o_vi	68.44	62.12	64.73	65.57
PIA_w_ph_w/o_geom	98.31	91.56	96.77	96.49
PIA_w_ph_w/o_arc	98.64	95.99	96.68	97.02
PIA_w_ph	98.95	92.54	96.43	96.66
PIA_w/o_EB0	94.81	81.70	86.54	88.68
Vgg16_w/o_PIA	91.51	78.36	85.49	86.62
PIA	<b>99.24</b>	<b>96.47</b>	<b>97.76</b>	<b>98.06</b>

Table 3. Ablation analysis on DeepSpeak v2.0 test set based on AUC (%) scores. Global refers to the combined test set provided in the dataset. The best results are highlighted in bold.

performance. By replacing EfficientNet-B0 CNN backbone in training with frozen pretrained RESNET-18 embeddings, we see the AUC dropped by 4.43%-pts, 14.77%-pts, 11.22%-pts, and 9.38%-pts for Lip-sync, Face-swap, Avatar, and Global test subsets, respectively.

**Using Vgg16 CNN model (Vgg16\_w/o\_PIA)** In this experiment, we replace our PIA detection module with a simple VGG16 [45] CNN architecture to assess the significance of phonemes, visemes, geometric, and identity cues for deepfake detection. Replacing our architecture with a simpler model resulted in the AUC dropping by 7.73%-pts, 18.11%-pts, 12.27%-pts, and 11.44%-pts for Lip-sync, Face-swap, Avatar, and the Global test subset, respectively.

From this analysis, we observe that integrating visual, geometric, and identity cues extracted using phonemes with an EfficientNet-B0 CNN backbone provides the most robust model to detect deepfakes.

## 6. Conclusion

In this paper, we offer PIA (Phoneme-Temporal and Identity-Dynamic Analysis), an innovative, unified multi-modal technique for audio-visual deepfake detection. PIA

concurrently models phoneme articulation, visual features, geometric consistency of lips, and identity indicators to identify subtle temporal and cross-modal inconsistencies. Our approach attains state-of-the-art performance, exhibiting robust generalisation in cross-manipulation settings. PIA demonstrates exceptional performance on the high-resolution DeepSpeak v2.0 dataset, demonstrating its resilience in authentic and high-fidelity deepfake contexts.

Although our method achieves strong results on videos at the specific resolutions used for training, its generalization remains limited to those conditions, additional fine-tuning is needed to perform reliably on videos with different resolutions. In addition, due to our reliance on WhisperX and wav2vec2 for phonetic alignment, the model has been restricted to English-language inputs. Lastly, our approach is designed for scenarios involving fake videos and may not be applicable to instances of RealVideo-FakeAudio (RVFA), where the visual elements are authentic but the audio has been manipulated.

The possible areas of extension to our work include generalization capabilities of our model to video resolutions beyond the training data, by augmenting the training data with varied resolutions. Another significant direction is addressing RealVideo-FakeAudio (RVFA) cases. We intend to extend the model’s capability to detect audio-only forgeries as well by introducing modality-specific anomaly detectors. Furthermore, to enhance multilingual deepfake detection, we intend to substitute the English-centric phonetic alignment with multilingual voice representations, such as those obtained by multilingual automatic speech recognition (ASR) speech-to-text models.

**Acknowledgment.** This work is supported by the Center for Identification Technology Research (CITeR) and the National Science Foundation under Grant No. 1822190.



## References

- [1] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020. 1, 3
- [2] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023. 3, 4, 6
- [3] Sarah Barrington, Matyas Bohacek, and Hany Farid. Deepspk dataset v1. 0. *arXiv preprint arXiv:2408.05366*, 2024. 6, 7
- [4] Mathieu Bernard and Hadrien Titeux. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958, 2021. 4
- [5] Matt Bracken. Cyber firm knowbe4 hired a fake it worker from north korea. <https://cyberscoop.com/cyber-firm-knowbe4-hired-a-fake-it-worker-from-north-korea/>. Published by CyberScoop; accessed 1-January-2025. 1
- [6] Heather Chen and Kathleen Magramo. Finance worker pays out \$25 million after video call with deepfake ‘chief financial officer’. <https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>. Published by CNN; accessed 30-January-2025. 1
- [7] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 2
- [8] Yu Chen, Yang Yu, Rongrong Ni, Yao Zhao, and Hao-liang Li. Npvoice: Jointing non-critical phonemes and visemes for deepfake detection. *arXiv preprint arXiv:2306.06885*, 2023. 3
- [9] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–22, 2023. 7
- [10] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2
- [11] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 439–447, 2020. 7
- [12] Soumya Kanti Datta, Shan Jia, and Siwei Lyu. Exposing lip-syncing deepfakes from mouth inconsistencies. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 3
- [13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 3, 4
- [14] Jiankang Deng, Jia Guo, Xiang An, Jack Yu, Baris Gecer, and Stefanos Zafeiriou. InsightFace: State-of-the-art 2d and 3d face analysis toolbox. <https://github.com/deepinsight/insightface>, 2023. GitHub repository, latest release 2023-04-02, accessed 2025-06-29. 4
- [15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [17] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10491–10503, 2023. 3, 8
- [18] FFmpeg Developers. Ffmpeg. <https://ffmpeg.org/>. Version 4.4.2. 4
- [19] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1
- [20] 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
- [21] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don’t lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 3, 6, 7, 8
- [22] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14950–14962, 2022. 7, 8
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [25] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023. 3
- [26] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Applied Soft Computing*, 136:110124, 2023. 7
- [27] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021. 6, 7

- [28] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 2, 6
- [29] Jonas Krause, Andrei De Souza Inacio, and Heitor Silvério Lopes. Language-focused deepfake detection using phonemes, mouth movements, and video features. In *2023 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. IEEE, 2023. 1
- [30] Chunyu Li, Chao Zhang, Weikai Xu, Jinghui Xie, Weiguo Feng, Bingyue Peng, and Weiwei Xing. Latentsync: Audio conditioned latent diffusion models for lip sync. *arXiv preprint arXiv:2412.09262*, 2024. 2
- [31] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 3
- [32] 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
- [33] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuoling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3, 4
- [34] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976. 2
- [35] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don’t lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020. 7
- [36] Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5292–5302, 2024. 2
- [37] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 2, 6
- [38] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6111–6121, 2021. 3
- [39] Trevine Oorloff, Surya Koppiseti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024. 6, 7, 8
- [40] Trevine Oorloff, Surya Koppiseti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024. 3
- [41] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 2, 6
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [43] 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. 3, 7, 8
- [44] Henry Ruhs. Facefusion, 2024. 2
- [45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 8
- [46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4, 7
- [47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 4
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5
- [49] Haofan Wang. Inswapper: Face swapping model based on insightface, 2023. 2
- [50] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*, 2021. 7
- [51] Qiantong Xu, Alexei Baevski, and Michael Auli. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*, 2021. <https://arxiv.org/abs/2109.11680>. 4
- [52] Wenyan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023. 3, 7
- [53] Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of clip in generalizable deepfake detection. *arXiv preprint arXiv:2503.19683*, 2025. 7
- [54] Cai Yu, Shan Jia, Xiaomeng Fu, Jin Liu, Jiahe Tian, Jiao Dai, Xi Wang, Siwei Lyu, and Jizhong Han. Explicit correlation learning for generalizable cross-modal deepfake detection. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 3

- [55] Shengkai Zhang, Nianhong Jiao, Tian Li, Chaojie Yang, Chenhui Xue, Boya Niu, and Jun Gao. Hellomeme: Integrating spatial knitting attentions to embed high-level and fidelity-rich conditions in diffusion models. *arXiv preprint arXiv:2410.22901*, 2024. [2](#)
- [56] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*, 2024. [2](#)
- [57] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. [3](#), [7](#), [8](#)
- [58] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14800–14809, 2021. [8](#)