PLAYING FOR YOU: TEXT PROMPT-GUIDED JOINT AUDIO-VISUAL GENERATION FOR NARRATING FACES USING MULTI-ENTANGLED LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a novel approach for generating realistic speaking and taking faces by synthesizing a person's voice and facial movements from a static image, a voice profile, and a target text. The model encodes the prompt/driving text, a driving image, and the voice profile of an individual and then combines them to pass them to the multi-entangled latent space to foster key-value and query for audio and video modality generation pipeline. The multi-entangled latent space is responsible for establishing the spatiotemporal person-specific features between the modalities. Further, entangled features are passed to the respective decoder of each modality for output audio and video generation. Our experiments and analysis through standard metrics showcase the effectiveness of our model. All model checkpoints, code and the proposed dataset can be found at: https://github.com/Playing-for-you.

1 INTRODUCTION

AI-generated real-time audio-video multimedia communication by rendering realistic human talking faces has recently drawn massive attention[1,2]. Such technology is promising in various applications such as digital communication, helping communicate with individuals with impairments, in designing artificial instructors, and in designing interactive healthcare (Xu et al., 2024a; Gan et al., 2023). In such applications, generating realistic and real-time speech and visual content simultaneously is a key requirement. Therefore, an ideal scenario would be given a prompt text along with a face image and the audio profile of an individual, a talking human face is screenplayed as output with audio (generated speech) and visual narrating according to the prompt text.

Generative AI has emerged as a key area of interest in the computer vision and learning representation community. While existing approaches have made significant strides, however, they are constrained by their reliance on generating single modality (Egger et al., 2020; Kim et al., 2021). For example, current text-to-speech models (TTSM) (Ao et al., 2022; Betker, 2022; Casanova et al., 2024) focus primarily on voice synthesis. Similarly, considering the visual generation techniques i.e talking faces models (TFM) (Ren et al., 2021; Rombach et al., 2022; Siarohin et al., 2020; Zhang et al., 2023a; Xu et al., 2024a;b; Zhang et al., 2023b) aims at face video generation given a text or/and audio or/and image as a prompt. Hence both TTSM and TFM techniques are not suitable for real-life audio-video multimedia communication scenarios such as audio-visual chatbots, as in such situations both realistic video and speech must be generated synchronously and simultaneously. Few efforts have been made in the literature to merge TTSM and TFM by cascading the pipeline (Wang et al., 2023; Zhang et al., 2022). Further, in (Jang et al., 2024) made an effort to generate talking face and speaking audio jointly for a specific individual from a prompt text.

Further, these TFM (Chen et al., 2024; Zhang et al., 2019) depend on guidance by the defined facial properties from the weakly supervised latent information from the reference modality. As a result poor lip-synchronization and limited to tuning an existing audio profile to personalize the video

[1] https://www.business-standard.com/technology/tech-news/odisha-television-introduces-lisa-india-s-first-ai-news-presenter-123071000767_1.html

[2] https://www.indiatoday.in/india/story/india-today-groups-ai-anchor-sana-wins-global-media-award-for-ai-led-newsroom-transformation-2532514-2024-04-27

content and as a result the generation is far from being realistic. Moreover, expressiveness concerning facial dynamics along with subtle nuances for realistic facial behaviour needs to simultaneously match with audio content temporally to produce realistic talking faces. Further, such synchronization also depends on individual traits, such as how an individual speaks as per their voice intonation and other covariates. Although they are supposed to be important considerations for realistic speaking and taking faces models (STFM), however, this was not in the scope of only work on STFM (Jang et al., 2024). Therefore, this gap in the literature motivates us to design a prompt text-guided audio-visual multimodal generative STFM that can jointly generate audio and video, given a reference image and reference audio along with the prompt text as input.

Consequently, deviating from the literature (See Figure 1), in this work, we introduce a novel multi-modal framework designed to address these limitations by generating highly realistic speech and animations from a combination of prompt text, a driving image, and an audio profile as inputs. To illustrate, our framework aims to synthesize videos of a talking human face where the person in the image appears to speak along with the generated voice from the provided text for the given identity. Our method enhances the capabilities of existing pre-trained models (Xu et al., 2024a) by an advanced parallel mechanism that leverages both visual and auditory data streams. This parallelism ensures that the synthesized videos not only align the subject's facial movements with the spoken text but also synchronize with the generated personalized voice outputs that correspond to the subject's appearance.

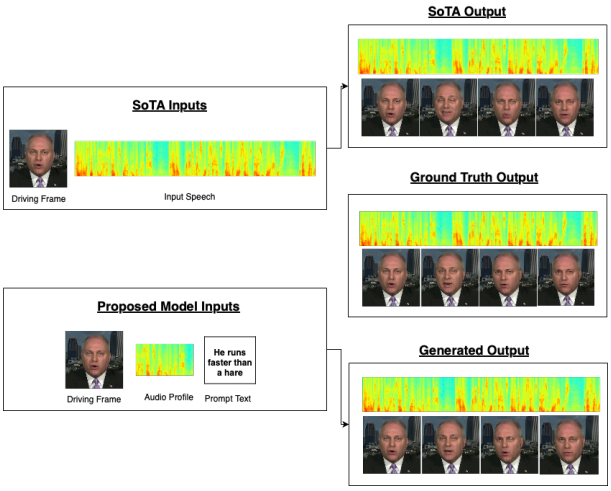


Figure 1: SOTA of taking faces were a driving frame as face image with an audio prompt is passed as input to the existing model such as Hallo (Xu et al., 2024a), VASA (Xu et al., 2024b) and the proposed model which generates a realistic audio-video synchronous multimodal taking faces with face image and audio profile of an individual along with the prompt text.

A person-agnostic generalized STFM model will encompass a large appearance and acoustic features variation. Furthermore, extracting such structure information along with the temporal synergy between the audio and video persevering personal variance will require additional modules to model these complexities. Hence, we introduce a parallel multiple entanglement on the latent space between the modalities' encoding and decoding engineering. The encoding stage develops the need for structure representation.

Our proposed architecture for STFM contains three main phases (See Figure 2). *Modality encoding phase*, at this stage a heterogenous personal signature of the audio and video modality, and the driving feature from the text are featured. The second stage is the *multi-entangled latent space* which glens the spatiotemporal relation and synchronisation in the embedding of the modalities, which further acts as the input to the *decoders phase* i.e the third stage of the proposed architecture. In the second stage, the exchange of information between the key and values i.e. the identity information (from audio and video extracted from the individual encoder) and queries (driving feature from encoded prompt text) are streamlined. To instrument this, an entanglement of the audio and text latent is performed which further entangles with video latent in transformers block and then to a diffusion block. The output of the diffusion block is passed to the video decoder. Similarly, an entanglement of the video and text latent is performed which further entangles with audio latent in a transformer space and passes to a text decoder block and then to the audio decoder. Such entanglements ensure to streamlined the audio profile and the driving image by linear navigation in the latent space along with the encoded feature from the prompt text. Specifically, the temporal information for both the audio and video generation is constructed by linear displacement of codes in the latent space as per the encoded text prompt. In turn, the model also learns a set of orthogonal motion directions to simultaneously learn the audio

and video temporal synergy, by exchanging their linear combination to represent any displacement in the latent space. To summarise our specific contributions are as follows:

- To the best of our knowledge, the proposed architecture is the first person-agnostic STFM which fosters a text-driven multimodal realistic audio-video synthesis that can generalised to any identity.
- We design a three-phase architecture which consists of the encoder, muti-entangled latent and decoder phase for audio and video pipeline. The muti-entangled latent space glens the spatiotemporal and synchronisation in the encoder embedding to exchange information between the modality and guided text and help for generating crucial visual and acoustic characteristics based on input profiles.
- With the comprehensive experiments, we demonstrate that the proposed method surpasses the state-of-the-art techniques available for STFM.

2 Related Work

Text-to-Speech, as a technology, has seen remarkable progress in recent years, with the development of models that generate highly natural and expressive speech. Modern Text-To-Speech approaches(Casanova et al., 2024; Betker, 2022) leverage sequence-to-sequence architectures to map text directly to speech. Notable ones among these are the Tacitron(Wang et al., 2017) and the newer Tacitron2(Shen et al., 2018) Models. These models employ attention mechanisms to convert text sequences into mel-spectrograms. These spectrograms are then passed through neural vocoders like WaveNet(van den Oord et al., 2016) or HiFi-GAN(Kong et al., 2020) to generate high-quality audio waveforms. Other models, such as FastSpeech(Ren et al., 2019) and VITS(Kim et al., 2021), introduce optimizations that improve the speed of speech generation while maintaining or enhancing the naturalness and clarity of the output. While models have advanced into more complex architectures, the baseline idea behind the speech generation remains the same. TortoiseTTS(Betker, 2022) is a modern, expressive TTS system with impressive voice cloning capabilities. This model incorporates a combination of Auto-Regressive Model, followed by a Diffusion Model(Ho et al., 2020), to convert the input text into mel-spectrogram frames, via discrete acoustic tokens. This model also follows the standard of a vocoder(Univnet)(Jang et al., 2021) for generating the audio from the spectrogram frames. Only a few works have been made in the literature to attend STFM by cascading the pipeline (Wang et al., 2023; Zhang et al., 2022). In (Jang et al., 2024) advancement is made by generating talking face and speaking audio jointly for a specific individual from a prompt text.

2.1 Face Reenactment and Lip-Sync Models

Recent advancements in face reenactment have enabled realistic video generation by synthesizing facial movements driven by audio inputs. Early models like SyncNet(Raina & Arora, 2022) focused on lip synchronization through facial key points and phoneme mapping but struggled with capturing detailed expressions and diverse facial structures. More recent models, such as LipGAN(K R et al., 2019) and Wav2Lip(Prajwal et al., 2020a), leverage GANs to improve lip-sync accuracy and generate more natural facial animations.

The synthesis of multimodal human videos, combining text, audio, and visual inputs, has advanced considerably in recent years. Early approaches focused on audio-driven models that primarily addressed lip-syncing, mapping speech inputs to corresponding facial movements. Models like SyncNet(Raina & Arora, 2022) played a crucial role in establishing baseline synchronization between audio and lip movements. However, these models often lacked expressive, natural face dynamics.

2.2 Diffusion-based Lip-Sync Models

Recent models have extended beyond simple lip-syncing to incorporate emotional expression and natural head motion. Audio2Head(Wang et al., 2021), for example, shifts from keypoint-based methods to a dense mapping of audio features onto facial expressions and head motion, resulting in a more fluid and expressive representation of speech-driven animations. Expressive Audio-driven Talking-heads (EAT)(Gan et al., 2023) enhances this by integrating text and audio as inputs, introducing more dynamic and natural facial expressions synchronized with speech.

The Hallo(Xu et al., 2024a) model builds on these advancements by using attention mechanisms to improve facial reenactment, ensuring smoother transitions and better coherence across diverse speakers. Furthermore, SadTalker(Zhang et al., 2023b) incorporates 3D facial representations,

combining both speech and facial dynamics for more realistic head motions and expressive gestures. FaceChain-ImagineID(**?**) uses Latent Diffusion to generate talking faces directly using only audio input, generating synthetic faces after disentangling the audio to extract aspects like expression, identity and emotion.

Other notable works, such as Diffused Heads(Stypułkowski et al., 2023) and DreamTalk(Zhang et al., 2023a), have explored diffusion-based models for video generation, leveraging the success of image-to-video transformations in generating high-quality talking-head videos. These models focus on temporally consistent video generation, addressing fidelity and synchronization across frames.

3 METHODOLOGY

Our proposed joint learning methodology for the audio, video, and natural language-based text prompts comprises three components – namely, **(1)** Encoding phase, **(2)** Entanglement of combined latent space, and **(3)** Decoding phase *i.e.,* Latent conditional generation of synthesized audio-video. Figure 2 illustrates detailed network architecture and roles of different model components to learn and dynamically synthesize audio-video on a given source image.

3.1 MULTI-MODAL ENCODING PHASE.

HiFi-GAN (Kong et al., 2020) and Wav2Vec Encoder (Baevski et al., 2020) are employed to extract high-dimensional embedding vectors from the reference audio. The HiFi-GAN generates feature embedding $\mathbf{f_a}$ representing the audio waveform. At the same time, the Wav2Vec encoder produces a secondary set of embedding $\mathbf{f_s}$ capturing semantic audio information. We assume that the semantic audio embedding is a direct mapping from the speaker's profile. Consequently, the combined features $\mathbf{f_a} \oplus \mathbf{f_s}$ provide a detailed audio profile necessary for driving the lip-sync and facial animations in the synthesized video. The input reference audio is represented as a 2-second MEL-spectrogram, *i.e.,* a sequence of acoustic features per frame of 0.2 seconds duration with the shape of $\mathbb{R}^{5609\times512}$.

Our neural model's newly inducted input text prompt undergoes Byte-Pair Encoding (BPE) and Tokenization (Zouhar et al., 2024) to convert textual information into a feature vector $\mathbf{f_t} \in \mathbb{R}^{512.T}$. This feature vector enables context-specific animations, allowing the synthesized video to align with the intended spoken words and expressions implied in the text. The purpose of concatenating $\mathbf{f_t}$ with the combined feature of reference audio $\mathbf{f_a} \oplus \mathbf{f_s}$ is to obtain the speaker's signature in the final flattened feature tokens of $\mathbf{f_t} \oplus \mathbf{f_a} \oplus \mathbf{f_s} \in \mathbb{R}^{5609+T\times512}$.

Next, the input source image is processed through a Variational Auto-Encoder (VAE) (Kingma & Welling, 2022) and a Landmarks Detection model (Zhang et al., 2020). The VAE generates an image embedding $\mathbf{f_i}$, representing the visual style and identity of the person in the source image. Concurrently, the landmarks detection network extracts structural features – face mask feature $\mathbf{f_{fm}}$ and lip mask feature $\mathbf{f_{lm}}$, which are combined with the image embedding vectors to create a fused visual feature representation $\mathbf{f_i} \oplus \mathbf{f_{lm}} \oplus \mathbf{f_{fm}} \in \mathbb{R}^{3136\times512}$. The straightforward tendency of traditional methods is either to introduce prior 3D morphable models faces (Zhang et al., 2023b), motion priors of the facial parts (Jang et al., 2024), or guiding video frames (Wang et al., 2022) to learn nuances of facial articulation in relation to the audio in combined latent space. In contrast, we derive how the entanglement of multiple latent spaces of text-audio-video using Transformer encoders (Vaswani et al., 2023) can eliminate the dependency on strong motion priors. As a result, we are able to use text prompt features as a set of anchoring tokens to both the Transformer encoders.

3.2 ENTANGLEMENT OF COMBINED TEXT-AUDIO-VIDEO LATENT SPACE.

As illustrated in the Figure. 2, a smooth synergy between the text-audio latent embedding and the text-image latent embedding is established by two Transformer encoders followed by latent diffusion-guided (Xu et al., 2024a) synthesizer of visual nuances and decoder-only GPT-2 (Casanova et al., 2024) model for synthesizing text-conditioned audio latent.

The first Transformer encoder spatially contextualizes the audio MEL-spectrogram tokens using a dual-stream cross-modal attention mechanism with the flattened version, denoted by $\mathbf{L(.)}$, of *categorically fixed speaker* embedding tokens merged with varying text embedding tokens, *i.e.,* $\mathbf{Q_a} = \mathbf{L}(\mathbf{f_a} \oplus \mathbf{f_s})$, as

$$\text{Cross-Attention}(\mathbf{Q_a}, \mathbf{K_{ti}}, \mathbf{V_{ti}}) = \text{SoftMax}\left(\frac{\mathbf{Q_a}\mathbf{K_{ti}}^{\top}}{\sqrt{d_k}}\right)\mathbf{V_{ti}}, \tag{1}$$

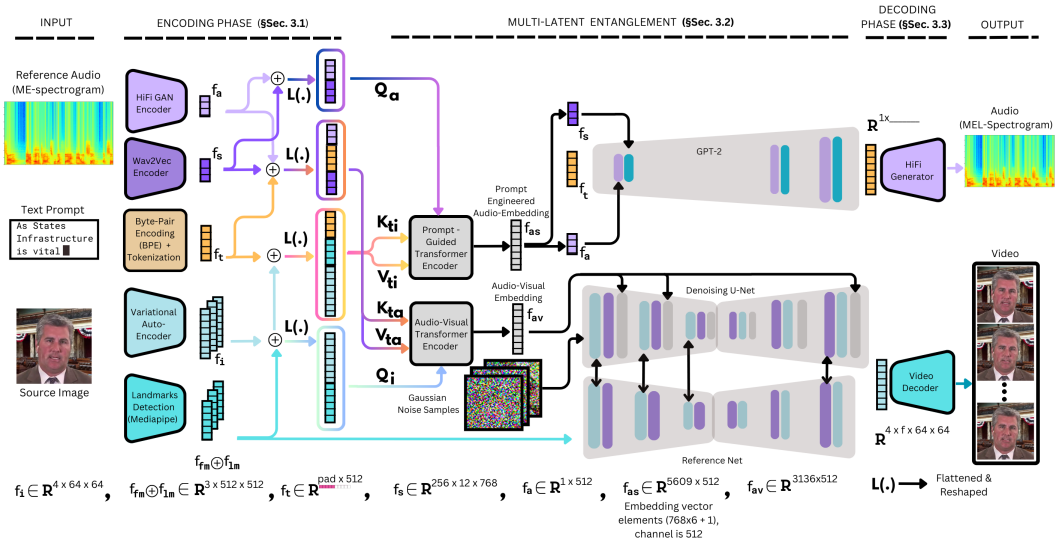


Figure 2: **Our Network Architecture:** Text Prompt-guided joint audio-visual learning representations using dual stream Transformer Encoders and Denoising Diffusion model. The model architecture can be divided into three phases – namely *Encoding Phase*, *Multi-Latent Entanglement*, and *Decoding Phase*. As an output, an audio-visual animation is generated from a single source image, reference audio, and a short text prompt.

where the query vector $\mathbf{Q}_a$ is of dimension $\mathbb{R}^{5609\times512}$ and the key-value paring $(\mathbf{K}_{ti}, \mathbf{V}_{ti})$ between the tokens of $\mathbf{L}(\mathbf{f}_t \oplus \mathbf{f}_i \oplus \mathbf{f}_{lm} \oplus \mathbf{f}_{fm})$ has a variable spatial length (padded up-to a max length) with a fixed channel length of 512. We assume that merging the varying text tokens has fulfilled two purposes – **(1)** first, querying audio tokens as well as the speaker tokens has been implicitly prompt-engineered by the text tokens, **(2)** second, when the resulting prompt-engineered latent embedding vectors $\mathbf{f}_{as}$ are split into its respective constituents, they become proxy weights of text-image embedding vectors.

Similar to the previous encoder block, the second Transformer encoder spatially contextualizes the input masked-image embedding vectors $\mathbf{L}(\boldsymbol{f}_i \oplus \boldsymbol{f}_{fm} \oplus \boldsymbol{f}_{lm})$ using cross-modal attention with the key-value pairs $(\mathbf{K}_{ta}, \mathbf{V}_{ta})$ of merged text-audio embedding tokens $\mathbf{L}(\boldsymbol{f}_t \oplus \boldsymbol{f}_a \oplus \boldsymbol{f}_s)$ similar to the equation 1 as

$$\text{Cross-Attention}(\mathbf{Q}_i, \mathbf{K}_{ta}, \mathbf{V}_{ta}) = \text{SoftMax}\left(\frac{\mathbf{Q}_i \mathbf{K}_{ta}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{ta}. \tag{2}$$

As a result, the output latent embedding on audio-visual features $\boldsymbol{f}_{av}$ can serve as a compact and compressed representation of facial animation sequences in the high-dimensional space. Therefore, our next step is to learn a synthesizer *i.e.,* a hierarchical latent diffusion model Xu et al. (2024a) for video generation and a corresponding MEL-spectrogram synthesizer based on the X-Text-to-Speech (XTTS) model Casanova et al. (2024).

Latent Text Conditioned Spectrogram Systhesizer: The GPT-2 encoder is based on the TTS model (Casanova et al., 2023) and (Shen et al., 2018). This part is composed of a decoder-only transformer module that is conditioned by the audio and speaker embedding vectors $\boldsymbol{f}_a, \boldsymbol{f}_s$ disentangled from the prompt-engineered audio embedding vector $\boldsymbol{f}_{av}$, and the auto-regressive generation of spectrogram tokens is fully driven by the input text tokens from $\boldsymbol{f}_{av}$.

Text-Anchored Audio-Video Latent Conditioned Denoising Diffusion: The Denoising Diffusion model aims to reverse a diffusion process(Ho et al., 2020; Song et al., 2022) that progressively adds random Gaussian noise noise to data. Inspired by the Hallo method (Xu et al., 2024a), we employ an additional augmentation of the text-anchored latent embedding vector learned to combine the audio and motion nuances on a single image inside the Denoising U-Net (Ronneberger et al., 2015) model of Hallo. The model is initialized with pre-trained weights and fine-tuned during the training step.

Throughout the diffusion processes, we introduce embedding cross-attention, which incorporates the combined latent space embedding, particularly our $\boldsymbol{f}_{av}$, into each diffusion step. This cross-attention mechanism allows the diffusion models to leverage the shared information across modalities, ensuring

that the generated outputs (audio and video) are consistent with the input embedding. The inclusion of cross-attention helps to maintain coherence between the synthesized motion across all the pixels of the source image.

Additionally, diffusion cross-attention facilitates mutual information exchange between the audio and video diffusion blocks. This cross-attention mechanism enables the audio and video models to synchronize their outputs, ensuring that the generated audio and video components are temporally aligned. By integrating this cross-attention, our framework effectively coordinates the diffusion processes, leading to synchronized and coherent multimedia output.

3.3 Decoding Phase for Audio-Video Generation

The outputs of the previous steps are processed by their respective final decoders. For audio generation, similar to the XTTS method (Casanova et al., 2024), the synthesized spectrogram is passed through a Vocoder component of HiFi Generator module to obtain the final audio signal. For video, the Denoising UNet-generated f number of frames of dimension $\mathbb{R}^{4\times f\times 64\times 64}$ are decoded by a pre-trained decoder component of (Kingma & Welling, 2019) to produce the complete video.

3.4 Loss Functions

To train our model, we use – **(1)** video loss at the pixel-level *i.e.,* sum of the N number of pixel intensities between the ground truth image frame $\mathcal{I}^f_{\text{gt}}$ and the generated frame $\mathcal{I}^f_{\text{gen}}$ for f number of frames as $\mathcal{L}_{\text{video}} = \sum_f \sum_{i=1}^{N} \|(\mathcal{I}^f_{\text{gt}})^i - (\mathcal{I}^f_{\text{gen}})^i\|$, **(2)** audio loss at the spectrogram $\mathcal{S}$ domain as mean squared error between the ground-truth magnitudes and generated magnitudes at different of time step t as $\mathcal{S}^t_{\text{gt}}$ and the generated frame $\mathcal{S}^t_{\text{gen}}$ as $\mathcal{L}_{\text{audio}} = \frac{1}{T}\sum_{t\in T} \|(\mathcal{I}^f_{\text{gt}})^i - (\mathcal{I}^f_{\text{gen}})^i\|^2$, and finally the total loss **(3)** as $\lambda\mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{video}}$ with balancing factor $\lambda = 0.1$.

4 Experimental Results

4.1 Datasets, Preprocessing, Implementation Details and Evaluation Matrices

Datasets: We have primarily conducted our experiments on 4 datasets. Our model training was done on a combination of **VoxCeleb** Dataset (Nagrani et al., 2019), **FakeAVCeleb** dataset (Khalid et al., 2022), **HDTF** (Zhang et al., 2021) and the **CelebV-HQ** dataset (Zhu et al., 2022). VoxCeleb is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. FakeAVCeleb is a novel audio-video multimodal deepfake dataset, we only considered the original part of the dataset. CelebV-HQ is a large-scale video facial attributes dataset, that demonstrates a diverse quality of data, which is important to test the robustness of our model for the application scenario for realistic speaking and talking face generation. HDTF is a large in-the-wild high resolution audio-visual dataset built for talking face generation.

Preprocessing: Our preprocessing involved resizing the videos to 512x512, and then cropping them down to 20 seconds by taking the 1st 20 seconds from each video sample (at 25FPS which equates to 500 frames). We then seperated the audio from the video using ffmpeg, and then ran the OpenAI's Whisper (Radford et al., 2022) to transcript the audio speeches.

Implementation details: The optimizer used for our model is AdamW with a weight decay of 1e-4 and weight decay of 1e-2, and the scheduler has a step-wise learning rate with a step size of 1000 and gamma of 0.5. The weight decay regularizes the model, preventing any overfitting. We have used Nvidia 1xA6000s GPU for training each model, and the model inference requires 12GB of VRAM. The total parameter size of the model comes to 1,575,936. The model performs 5.39 GFLOPs(Giga Floating Point Operations) during generation. We have trained the models for 10 epochs, with a batch size of 8.

Evaluation Metrics: Following are the evaluation matrices employed. **Video Metrics**: *Fréchet Video Distance (FVD):* A measure of the quality of generated videos, comparing them to real videos based on spatio-temporal features. Lower values indicate better performance (Unterthiner et al., 2019). *FID (Fréchet Inception Distance):* Evaluates the visual quality of individual frames by comparing the distributions of generated and real images. Lower scores represent better visual quality (Heusel et al., 2018). *Fréchet Video Motion Distance(FVMD):* Measures the quality of motion in generated videos, capturing the difference between real and generated motion trajectories. Lower values signify a more realistic motion.(Liu et al., 2024).*Lip-Sync Error - Confidence*: Measures how well the lip movements align with the corresponding audio. Higher scores indicate better synchronization. **Audio**

Metrics: *Fréchet Audio Distance (FAD):* Assesses the similarity between generated and real audio samples, with lower scores indicating closer resemblance. *Short-Time Objective Intelligibility (STOI):* Measures the intelligibility of the generated speech. Higher values represent more intelligible speech (Kilgour et al., 2019). Mel Cepstral Distortion(MCD): A metric used to evaluate the quality of speech synthesis by comparing the spectral features of generated and reference audio. Lower scores imply better audio quality (Zezario et al., 2020). **Audio-visual (AV) synchronisation**: We used two metrics proposed in Wav2Lip Prajwal et al. (2020b) to find the audio-visual synchronisation. The first is the average error measure calculated in terms of the distance between the lip and audio representations, "LSE-D" ("Lip Sync Error Distance"). A lower LSE-D denotes a higher audio-visual match, i.e., the speech and lip movements are in synchronization. The second metric is the average confidence score, "LSE-C" (Lip Sync Error Confidence). The higher the confidence, the better the audio-video correlation.

Training and Testing: Our primary training dataset is the VoxCeleb dataset(Nagrani et al., 2019), where our training set comprised of approximately 36000 videos. We chose this training set by filtering out individuals whose speech was in English. We tested on more than 200 samples from each of the four datasets, resulting in a test set of over 800 unseen samples.

We benchmarked the video outputs for the unseen samples against SoTA Portrait Animation models, like Hallo(Xu et al., 2024a), Sadtalker(Zhang et al., 2023b), EAT(Gan et al., 2023) and Audio2Head(Wang et al., 2021).

We also benchmarked the audio outputs for the unseen samples against SoTA Speech generation models, like Tortoise(Betker, 2022), Your_TTS(Casanova et al., 2023), XTTS_v2(Casanova et al., 2024) and GlowTTS(Kim et al., 2020).

4.2 Result Analysis

Video Results: From Table 1, we can observe that our model shows superior performance across all three metrics FID, FVD, and FVMD on VoxCeleb, CelebV-Hq and HDTF. This indicates high fidelity and minimal discrepancies are attended by the proposed model. On the FakeAVCeleb, the performance is slightly poorer but can be comparable, it still maintains strong visual consistency and realism on visual inspection. For the CelebV-HQ our model excels again, demonstrating its capability to produce high-quality video outputs.

Table 1: Video pipeline evaluation scores across datasets.

Dataset	Model	FID Score (↓)	FVD Score (↓)	FVMD Value (↓)
VoxCeleb	Audio2Head	81.00	90.12	5100.92
	Hallo	67.28	70.69	5703.44
	EAT	85.16	80.38	4878.36
	SadTalker	119.36	112.77	6352.19
	Our Model	**42.88**	**49.78**	**4192.07**
FakeAVCeleb	Audio2Head	93.59	97.85	1329.23
	Hallo	**26.88**	**39.42**	2351.20
	EAT	94.34	98.49	**1324.91**
	SadTalker	81.77	77.10	4158.18
	Our Model	47.24	49.15	2263.54
CelebV-HQ	Audio2Head	90.22	102.76	2939.49
	Hallo	42.76	56.10	2816.68
	EAT	47.88	56.21	2894.31
	SadTalker	52.60	52.55	2789.19
	Our Model	**34.01**	**43.67**	**2743.29**
HDTF	Audio2Head	37.78	32.69	2633.04
	Hallo	20.54	25.81	**1290.57**
	EAT	29.57	29.34	2573.05
	SadTalker	22.34	23.57	2410.89
	Our Model	**11.72**	**15.58**	1784.16

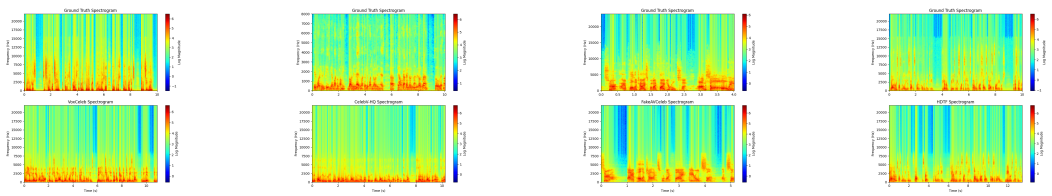
On the HDTF our model shows incredible performance in the FID and FVD metrics, beating all the other models, while our model is admirably performing considering FVMD when compared to Hallo. Therefore we observed that for some datasets, certain models work slightly better than the

Table 2: Audio pipeline evaluation scores across datasets.

Dataset	Model	FAD Score (↓)	MCD Score (↓)	STOI Score (↑)
VoxCeleb	Tortoise	258.54	82.37	0.10
	Your_TTS	**199.52**	111.79	**0.19**
	XTTS_v2	249.17	100.80	0.13
	GlowTTS	329.21	103.94	0.15
	Our Model	241.75	**75.39**	0.17
FakeAVCeleb	Tortoise	871.14	82.12	0.10
	Your_TTS	445.38	65.60	**0.21**
	XTTS_v2	184.39	77.88	0.11
	GlowTTS	482.04	87.11	0.18
	Our Model	**171.52**	**55.12**	0.19
CelebV-HQ	Tortoise	529.06	113.18	0.09
	Your_TTS	520.01	137.58	0.16
	XTTS_v2	509.90	124.61	0.07
	GlowTTS	549.18	139.81	**0.22**
	Our Model	**244.83**	**85.76**	0.18
HDTF	Tortoise	425.30	67.15	0.11
	Your_TTS	467.42	49.38	0.15
	XTTS_v2	135.11	49.65	0.14
	GlowTTS	510.61	66.42	0.12
	Our Model	**106.43**	**44.05**	**0.15**

Figure 3: The figures in each row show frames from the videos generated by each technique in the order: Ground Truth, Our proposed Model, Audio2Head (Wang et al., 2021), EAT (Gan et al., 2023), Hallo (Xu et al., 2024a), and SadTalker (Zhang et al., 2023b) on the VoxCeleb Dataset. A frame in each column for both videos corresponds to the same time-stamp (frames were sampled at equal intervals of 25 seconds across the videos).

proposed model, and the reason behind this is that those models try to memorise certain properties from individual datasets. Whereas our model is a more generalised version that can performed consistently on cross datasets having varying resolution, and video quality. The visualization from Figure 3 also concludes that our model can generate video very close to the ground truth and better than any model. From Figure 4 it can be concluded that our model can generate nearby results for HDTF, FakeAVCeleb and CelbV-HQ when compared to ground truth.



(a) (b) (c) (d)

Figure 5: Ground Truth vs. Generated Audio Spectrograms for (a) VoxCeleb, (b) CelebV-HQ, (c) FakeAVCeleb and (d) HDTF datasets

Audio Results: We can infer from Table 2 that our model consistently performs the best in the MCD Score metric, which suggests that it minimizes distortion between the spectral features of synthetic and reference speech. While considering the FAD scores, our model also performed on par state-of-the-art, except on VoxCeleb where Your_TTS is better, these showcase that the proposed model can generate consistently similar audio compared to the ground truth. Considering the STOI metric, the performance of our model is similar to or slightly lower than Your_TTS. The analysis of all the measures showcases that our model is more generalized and realistic as it can minimize distortion and also generate accurate distributions, and maintain intelligibility of the speech consistently better than any other models. The visualization from Figure 5 also concludes that our model can generate audio very close to the ground truth.

AV synchronization results: From Table 3 we can conclude that our proposed model has performed better audio-video synchronization than SOTA and is close to the ground truth. The proposed model has the lowest LSE-D, i.e. better audio-visual match, i.e. and LSE-C i.e. better audio-video correlation. We have also analyzed the model with varying accents, blurred audio profiles, and audio profiles of a kid with a source image of an adult and vice versa, and the results were found to be effective no bias was found in any aspect.

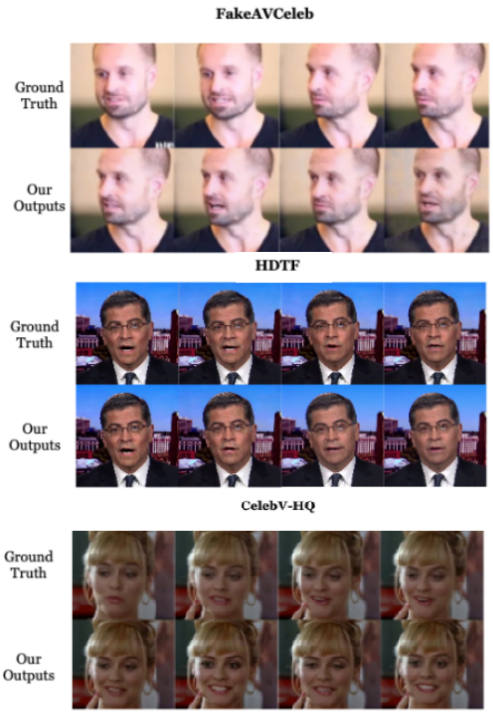


Figure 4: Results of our model on FakeAVCeleb, Celeb-HQ and HDTF datasets.

4.3 Ablation Study

Table 4 shows the ablation study of our proposed model. We have 3 main sub-networks that define the output of our model. The **Transformer Encoder Block(TE)** (Vaswani et al., 2023) with two variations shared-TE **(STE)** where both audio and video pipeline shares a transformer block and explicit-TE **(ETE)** where audio and video pipeline has explicit or separate transformer block. **Diffusion(Song et al., 2022) Cross Attention(DC)**, and the **Embedding Cross Attention(EC)**. From our results, it is understandably explainable that the transformer encoder block, which encodes our inputs

Table 3: Evaluation of audio-visual synchronization

	LSE-C(↑)	LSE-D(↓)
Groundtruth	5.45	8.52
Explicit	5.71	8.41
Hallo	3.03	8.71
Audio2Head	2.51	10.34
EAT	4.39	9.35
SadTalkert	5.44	10.09
Proposed	5.74	8.38

into a common latent space, is the most important
modality of our network, with its removal drastically reducing our metric values. Our experiments also show that the cross-attention blocks between the diffusion models are more important than the embedding cross-attention since our metric values drop more when we remove the diffusion cross-attention, probably since the diffusion cross-attention already syncs the modalities during the parallel learning stage. Another important aspect of ablation is the encoding latent in the individual transformer i.e. ETE is much better than STE. This infers that it is important to encode the latent for each modality separately while sharing information among the generated modalities. Table 5 shows our ablation study on the encoders. "Only Visual Tokens Attended" involves eliminating the audio prompt-guided transformer. Similarly, the "Only Audio Tokens Attended" involves using only the audio prompt-guided transformer encoder. "No Hifi-GAN" and "No Wav2Vec" are results obtained by eliminating the encoding process of the Hifi-GAN and Wav2Vec Models respectively. All this ablation concludes the importance of each of the components.

Table 4: Ablation study of the transformers.

ETE	STE	DC	EC	FID (↓)	FVD (↓)	FVMD (↓)	FAD Score (↓)	MCD (↓)	STOI (↑)
		✓	✓	86.70	80.88	5275.89	328.27	95.44	0.07
	✓			68.83	74.19	4412.74	260.91	87.51	0.11
	✓	✓		63.68	71.38	4298.30	250.12	83.96	0.14
	✓	✓	✓	61.44	69.15	2720.41	241.77	81.60	0.17
✓		✓	✓	**42.88**	**49.78**	**4192.07**	**241.75**	**75.39**	**0.17**

Table 5: Ablation study of the encoders.

Ablation	**FID (↓)**	**FVD (↓)**	**FVMD (↓)**	**FAD Score (↓)**	**MCD (↓)**	**STOI (↑)**
Only Visual Tokens Attended	68.31	78.42	5747.04	304.98	81.17	0.13
Only Audio Tokens Attended	69.02	79.35	6576.85	301.49	80.65	0.13
No Hifi-GAN	85.25	94.28	7483.40	498.33	87.51	0.09
No Wav2Vec	70.10	80.96	5926.64	309.95	89.58	0.11
Proposed Model	**42.88**	**49.78**	**4192.07**	**241.75**	**75.39**	**0.17**

4.4 Social Risks and Mitigations

There are social risks with technology development for text-driven by audio video talking face generation. The foremost risk is the ethical implications of creating highly realistic talking faces could be for malicious purposes, such as deepfakes. To mitigate such risk, ethical guidance for use of such generation techniques is required. Also, concerns regarding privacy and consent are implicit in such work. Transparent data usage policies by consent, and safeguarding the privacy of individuals can mitigate such concerns. By addressing these we aim to promote responsible and ethical generation technology.

5 Conclusion

This paper introduces a novel method for realistic speaking and talking faces by joint multimodal video and audio generation. We provide a holistic architecture where the information is exchanged between the modalities via the proposed multi-entangled latent space. A source image of an individual as a driving frame, reference audio which can be referred to as the audio profile of the individual and a driving or prompt text is passed as an input. The model encodes the input driving image, prompt/driving text, and the voice profile which are further combined and passed to the proposed multi-entangled latent space consisting of two separate transformers and diffusion block for video and text decoder for audio pipeline to foster key-vale and query representation for each modality. By this spatiotemporal person-specific featuring between the modalities is also established. The entangled-based learning representation is further passed to the respective decoder of audio and video modality for respective outputs. Conducted experiments and ablation studies prove that the proposed multi-entangled latent-based learning representation has helped our model obtain superior results on both video and audio outputs as compared to state-of-the-art models. While there is always scope for improvement in the future, we believe that our model has shown promising new learning representation for realistic speaking and talking face generation models.

REFERENCES

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing, 2022. URL https://arxiv.org/abs/2110.07205. 1

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477. 4

James Betker. Tortoise text-to-speech, 2022. URL https://github.com/neonbjb/tortoise-tts. Accessed: [date you accessed the repository]. 1, 3, 7

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone, 2023. URL https://arxiv.org/abs/2112.02418. 5, 7

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. Xtts: a massively multilingual zero-shot text-to-speech model, 2024. URL https://arxiv.org/abs/2406.04904. 1, 3, 4, 5, 6, 7

Ken Chen, Sachith Seneviratne, Wei Wang, Dongting Hu, Sanjay Saha, Md. Tarek Hasan, Sanka Rasnayaka, Tamasha Malepathirana, Mingming Gong, and Saman Halgamuge. Anifacediff: High-fidelity face reenactment via facial parametric conditioned diffusion models, 2024. URL https://arxiv.org/abs/2406.13272. 1

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future, 2020. URL https://arxiv.org/abs/1909.01815. 1

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation, 2023. URL https://arxiv.org/abs/2309.04946. 1, 3, 7, 8

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500. 6

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239. 3, 5

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation, 2021. URL https://arxiv.org/abs/2106.07889. 3

Youngjoon Jang, Ji-Hoon Kim, Junseok Ahn, Doyeop Kwak, Hong-Sun Yang, Yoon-Cheol Ju, Il-Hwan Kim, Byeong-Yeol Kim, and Joon Son Chung. Faces that speak: Jointly synthesising talking face and speech from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8818–8828, 2024. 1, 2, 3, 4

Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19. ACM, October 2019. doi: 10.1145/3343031.3351066. URL http://dx.doi.org/10.1145/3343031.3351066. 3

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022. URL https://arxiv.org/abs/2108.05080. 6

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019. URL https://arxiv.org/abs/1812.08466. 7

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search, 2020. URL https://arxiv.org/abs/2005.11129. 7

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, 2021. URL https://arxiv.org/abs/2106.06103. 1, 3

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056. 6

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114. 4

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020. URL https://arxiv.org/abs/2010.05646. 3, 4

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos, 2024. URL https://arxiv.org/abs/2407.16124. 6

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019. 6, 7

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, MM '20. ACM, October 2020a. doi: 10.1145/3394171.3413532. URL http://dx.doi.org/10.1145/3394171.3413532. 3

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*, pp. 484–492, 2020b. 7

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356. 6

Akshay Raina and Vipul Arora. Syncnet: Using causal convolutions and correlating objective for time delay estimation in audio signals, 2022. URL https://arxiv.org/abs/2203.14639. 3

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech, 2019. URL https://arxiv.org/abs/1905.09263. 3

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering, 2021. URL https://arxiv.org/abs/2109.08379. 1

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752. 1

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597. 5

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018. URL https://arxiv.org/abs/1712.05884. 3, 5

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation, 2020. URL https://arxiv.org/abs/2003.00196. 1

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502. 5, 9

Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation, 2023. URL https://arxiv.org/abs/2301.03396. 4

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric challenges, 2019. URL https://arxiv.org/abs/1812.01717. 6

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. URL https://arxiv.org/abs/1609.03499. 3

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762. 4, 9

Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion, 2021. URL https://arxiv.org/abs/2107.09293. 3, 7, 8

Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 4

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017. URL https://arxiv.org/abs/1703.10135. 3

Zhichao Wang, Mengyu Dai, and Keld Lundgaard. Text-to-video: a two-stage framework for zero-shot identity-agnostic talking-head generation. *arXiv preprint arXiv:2308.06457*, 2023. 1, 3

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, 2024a. URL https://arxiv.org/abs/2406.08801. 1, 2, 3, 4, 5, 7, 8

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time, 2024b. URL https://arxiv.org/abs/2404.10667. 1, 2

Ryandhimas E. Zezario, Szu-Wei Fu, Chiou-Shann Fuh, Yu Tsao, and Hsin-Min Wang. Stoi-net: A deep learning based non-intrusive speech intelligibility assessment model, 2020. URL https://arxiv.org/abs/2011.04292. 7

Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Xiaohu Guo, and Jiashi Feng. Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation, 2023a. URL https://arxiv.org/abs/2312.13578. 1, 4

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking, 2020. URL https://arxiv.org/abs/2006.10214. 4

Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2659–2663. IEEE, 2022. 1, 3

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation, 2023b. URL https://arxiv.org/abs/2211.12194. 1, 3, 4, 7, 8

Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. One-shot face reenactment, 2019. URL https://arxiv.org/abs/1908.03251. 1

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021. 6

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, 2022. URL https://arxiv.org/abs/2207.12393. 6

Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. A formal perspective on byte-pair encoding, 2024. URL https://arxiv.org/abs/2306.16837. 4