Diffusion Model for Text to Synchronized Joint Talking Head Video and Audio Generation

Zhenfei Zhang*, Tsung-Wei Huang†, Guan-Ming Su†, Ming-Ching Chang*, and Xin Li*
*University at Albany, SUNY
†Dolby Laboratories

***Abstract*—In this work, we propose to use text as input to jointly generate video and audio with lip synchronization, aiming at the ultra-low bit rate video conferencing problem and other generative taking head Audio/Video scenario. Our method focuses on employing the Denoising Diffusion Probabilistic Model (DDPM) [1] with a double-branches U-Net architecture to produce synchronized video-audio pairs from textual inputs. Through our approach, only low-bitrate texts are transmitted during the transmission process, allowing for efficient data transfer. Moreover, we provide users with multiple options to generate video content. They can create videos using video captions, speech text, or even opt for no-text input. Additionally, users have the flexibility to input static images to customize the style of the output video. By revolutionizing the way video and audio are generated and transmitted, our goal is to make a valuable contribution to enhancing the efficiency and versatility of multimedia applications.**

***Index Terms*—Denoising Diffusion Probabilistic Model (DDPM), Video conferencing, Talking head, Ultra-low bit rate**

I. Introduction

Recent advancements in artificial intelligence (AI) have spurred significant interest in media creation techniques, particularly in the domains of text-to-image [2]–[5] and text-to-video [6], [7] generation. While existing methods have demonstrated the capability to produce high-quality media outputs from textual inputs, they typically focus on generating a singular modality, which may not align with the multi-faceted demands of real-world applications. The exploration of techniques for generating multiple modalities concurrently remains an area ripe for investigation. Additionally, contemporary video conferencing software often imposes high bitrate requirements, leading to inefficiencies in both bandwidth usage and adaptability to diverse contexts.

In order to address these challenges, in this work, we propose a method that can generate synchronized video-audio pairs using Denoising Diffusion Probabilistic Model (DDPM) [1] from low bitrate textual inputs, as depicted in Fig.1. In this work, we consider two users: the producer, who intends to share video content either online or directly with the receiver, and the receiver, who seeks to receive the shared video. Our method transmits only low bitrates texts, thereby dramatically reducing the transmission costs compared to sending video directly. The producer can send video description texts and/or talking texts, or even nothing; the receiver can generate the synchronized video-audio pairs.

Fig. 2 illustrates the diverse options available to the receiver for generating synchronized video content. The input on the receiver side is text. Once the text is received, the user can explore multiple possibilities to convert it into video and audio content, including the following options: 1) They have the option to directly convert the text to video and audio without any modifications, as shown in Fig. 2(a), 2) Alternatively, they can alter the video caption to generate a different video, as depicted in Fig. 2(b), 3) Another possibility is to modify the talking text, resulting in a video with different talking content, as seen in Fig. 2(c), 4) Additionally, users can input a static image to change the head or style of the generated video, exemplified in Fig. 2(d). In addition, our text-to-speech conditional generation operates in fully zero-shot mode, enabling acceptance of any input texts for our model. This crucial capability makes our technology highly versatile and applicable in various daily life scenarios.

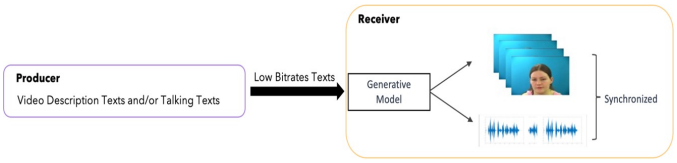

Fig. 1. The high-level pipeline of this project. In this setting, the generative model gets low bitrates text as inputs and generate synchronized video-audio pairs.

Our contributions can be succinctly summarized in the following aspects:

- We introduced a novel task of generating synchronized joint video and audio from textual inputs, effectively reducing transmission bitrates and enhancing efficiency.
- To ensure user flexibility, we provide multiple options for converting textual inputs into video-audio pairs, enabling personalized and tailored video creation experiences.

II. Related Work

A. Denoising Diffusion Probabilistic Model

The Denoising Diffusion Probabilistic Model (DDPM) [1] has garnered increasing attention owing to its formidable generative capabilities. It belongs to the category of generative models renowned for their adeptness at learning the underlying distribution of a training dataset and subsequently generating new instances. The DDPM operates through two principal processes: forward processing, which introduces noise, and backward processing, which reconstructs samples from pure

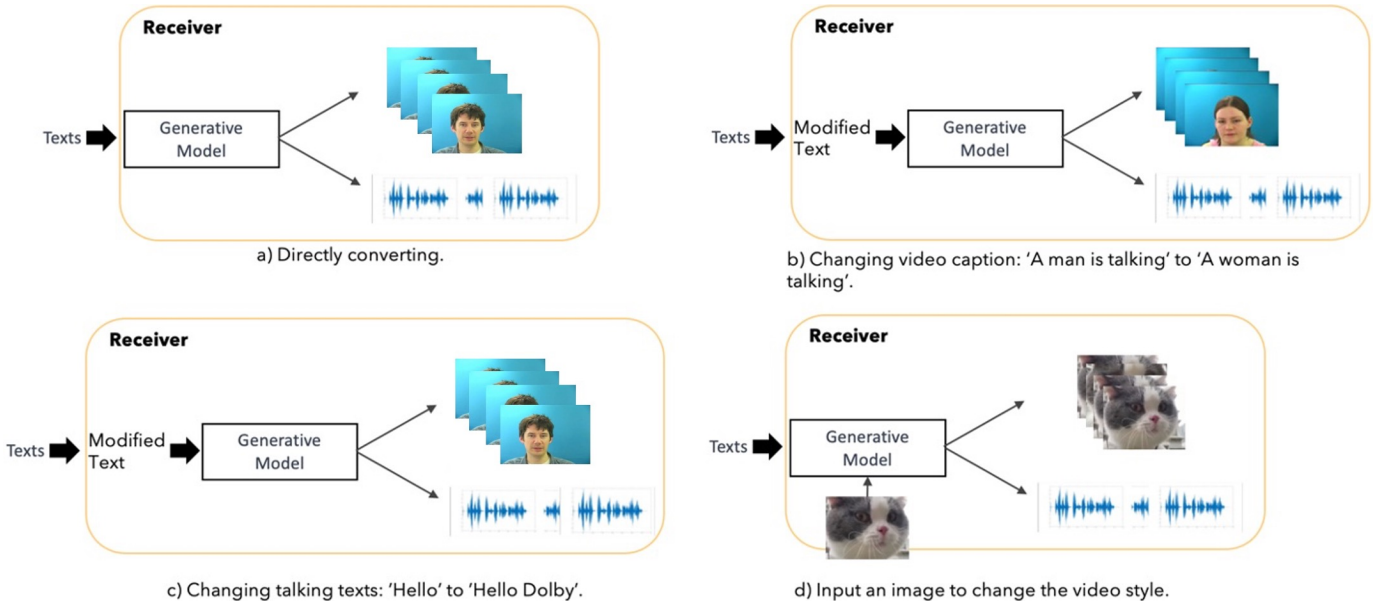

Fig. 2. Multiple options for the receiver to convert video-audio from texts.

noise. Notably, diffusion models [1] have demonstrated their proficiency in unconditional generation, relying solely on random distribution to generate novel instances that adhere to the learned characteristics of the training dataset. However, unconditional generation lacks controllability, rendering it unsuitable for real-world applications. Consequently, there has been a surge in research focusing on controllable generation through the integration of conditions, leading to the emergence of conditional diffusion models. These models incorporate various conditions, such as text [8]–[10], image [11], [12], audio [13], [14], among others, to guide the diffusion generation process. Beyond media creation tasks, diffusion models find application in diverse domains, including image segmentation [15], [16], change detection [19], [20], and image classification [17], [18], showcasing their versatility and excellence.

B. Cross-Modal Content Synthesis

The Cross-Modal Content Synthesis task encompasses the generation of diverse modalities from a given input modality, including text-to-image [2]–[5], text-to-video [6], [7], text-to-audio [21]–[23], among others. In addition to media creation from textual descriptions, recent efforts have been directed towards audio-driven talking-head video generation [24]–[26], aiming to produce videos synchronized with input audio, particularly focusing on achieving lip synchronization. However, many existing cross-modal content creation methods are constrained to generating only one modality at a time. In 2023, [13] introduced the first joint video-audio generation approach, although its output lacked controllability as it was generated from random distributions.

To overcome the limitations of previous methods and better align with real-world applications, we propose a novel approach: text-to-joint audio-video generation. Our method accepts two types of input texts that are video captions and speech content, and is capable of generating synchronized talking-head videos. This advancement represents a significant step forward in cross-modal content synthesis, offering enhanced flexibility and applicability for various multimedia generation tasks.

C. Pre-trained Models

To optimize training costs and elevate the quality of generated outputs, we leverage two pre-trained models: YourTTS [21] and Wav2Lip [23]. YourTTS facilitates the conversion of text into speech audio, while Wav2Lip enhances synchronization between video and audio components. Through the integration of these pre-trained models, we achieve efficient generation of high-quality results with enhanced coherence between video and audio elements. The utilization of pre-trained models is pivotal for enabling zero-shot generation from text to speech, and similar pre-trained models can be employed interchangeably for this purpose.

III. Preliminary of DDPM

In this section, we provide a brief overview of the Denoising Diffusion Probabilistic Model (DDPM) [1] and its fundamental processes. The DDPM consists of two primary procedures: forward processing and backward processing. We denote the original sample without noise and its corresponding pure Gaussian distribution as x_0 and x_T, respectively.

During forward processing, noise is incrementally added to the original sample x_0 over T time steps, gradually transforming it into pure Gaussian noise x_T. Conversely, backward processing aims to reconstruct the sample from the Gaussian distribution. The integration of these two processes enables the DDPM to generate high-quality samples while preserving the original data distribution. Due to its effectiveness, the DDPM has become a preferred choice in various generative tasks.

The primary objective of DDPM is to reconstruct x_0 from x_T. Therefore, the forward processing, following a Markov Chain [27] approach, can be represented as follows:

$$q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \tag{1}$$

$$q\left(x_1: x_T \mid x_0\right)=\prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \tag{2}$$

Where $t \in [0, T]$ is the time steps that is pre-defined, to gradually adding noise from true data x_0 to the pure Gaussian noise x_T. The diffusion forward process is used to get sample x_t from x_{t-1} by adding noise using normal distribution $\mathcal{N}(\cdot)$. $\sqrt{1-\beta_t}$ is the mean value and β is pre-defined variance. I is identity matrix. Using the property of forward pass, we can compute x_t from x_0 using:

$$x_t=\sqrt{\bar{\alpha}_t} x_0+\sqrt{1-\bar{\alpha}_t} \epsilon, \epsilon \sim \mathcal{N}(0,1) \tag{3}$$

Where $\alpha_t=1-\beta_t$ and $\bar{\alpha}_t=\prod_{i=1}^{t} \alpha_i$. And ϵ is the noise from normal distribution.

For the backward processing, the main goal is to train a model θ that can approximate $q(x_{t-1}|x_t, x_0)$ and $P_\theta(x_{t-1}|x_t)$. Therefore, we have the following formula for the reverse process:

$$p_\theta(x_{t-1}|x_t)=\boldsymbol{N}\left(x_{t-1} ; \mu_\theta(x_t, t) ; \sum_\theta(x_t, t)\right) \tag{4}$$

Where μ_θ is the mean value predicted from the trained model θ. Normally, we apply U-Net [28] as the model θ, and the training objective involves minimizing the loss function between the predicted noise and ground-truth noise, or between the predicted de-noised sample and the ground-truth sample. The training objective is:

$$U(\theta)=\mathbb{E}_{t, x_{0 \sim data}, \epsilon \sim(0, I)}\left[\|\epsilon-\epsilon_\theta(x_t, t)\|_2^2\right] \tag{5}$$

Where $\epsilon_\theta(x_t, t)$ is predicted noise from U-Net.

IV. Method

In this section, we will provide a detailed exploration of our methods, covering aspects such as the model architecture, unconditional generation, caption conditional generation, speech text conditional generation, and static images conditional generation. Our model demonstrates remarkable flexibility by accommodating various types of conditional inputs, including all, partial, or no conditional inputs. In scenarios where no speech text input is available, we input random distribution for audio generation, and the pre-trained YourTTS [21] is removed from the pipeline. Despite the absence of specific speech text, our model still generates dynamic video content by generating audio from random distribution, thereby aligning the audio and video components to produce synchronized output. It's worth noting that in the unconditional setting, the use of Wav2Lip [23] is unnecessary since the audio is generated from random distribution. Similarly, in the text conditional setting, the use of Wav2Lip is optional if the input speech text falls within the distribution of the training dataset, such as containing the same word or words with similar pronunciation. However, it is crucial for our model to be capable of accepting any speech text from the user.

A. Model Architecture

As mentioned earlier, the diffusion model typically utilizes the U-Net architecture as its foundational network structure. However, the original U-Net architecture is constrained to processing only one type of input, which doesn't fully align with the requirements of this project. To address this limitation, drawing inspiration from [13], we employ a double-branch U-Net capable of concurrently processing both video and audio inputs. The high-level pipeline is depicted in Fig.3(a). Unlike the original U-Net block, the double-branched U-Net block can handle two inputs simultaneously. Fig.3(b) and Fig.3(c) illustrate the video convolution block and audio convolution block, respectively, both incorporating group normalization and Swish activation functions [35]. Subsequently, the video feature and audio feature undergo separate convolutional layers to accommodate their distinct feature dimensions. Additionally, a cross-attention layer is introduced to facilitate learning the alignments between audio and video elements. We define the cross-attention function [29] as $A(\cdot)$, which can be succinctly represented as follows:

$$A(a_t, v_t)=Softmax\left(\frac{Q_a K_v^T}{\sqrt{d_k}}\right) V_v \tag{6}$$

$$K_v=linear(flatten(v_t)) \tag{7}$$

Where a_t and v_t are the video and audio in the t time step, respectively. Q, K and V are query, key and value. d_k denotes to the dimension of K_v. Similarly, $A(v_t, a_t)$ can be computed using symmetrical way the same as [13].

B. Unconditional Generation

The model underwent training using a dataset consisting of synchronized audio-video sentence pairs, enabling it to acquire unconditional generation capability. Consequently, the model is capable of generating video-audio content from random distributions, demonstrating its versatility and robustness in producing coherent multimedia outputs without specific input constraints.

C. Caption Conditional Generation

Caption conditional generation serves to control the generation process by providing captions as input. For instance, when we supply the caption "A man is talking" to an unconditional generative model, the model can produce video-audio content featuring a man's head. To realize captioned conditional generation, the same as [30], we offer two distinct approaches: classifier guidance and conditional generative diffusion model. Through the inclusion of video captions, the caption conditional model achieves fine-grained control over

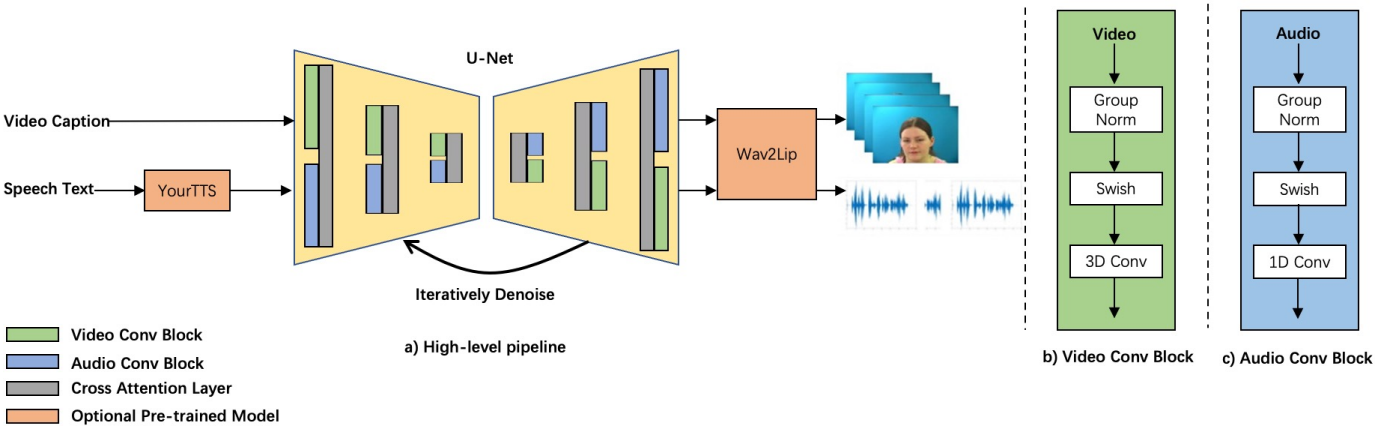

Fig. 3. a) The high-level pipeline of the method. b) and c) show the architecture of the video convlutional block and audio convolutional block, respectively. The components in orange are optional pre-trained model, which aim to enhance the user flexibility and generation quality.

the generated video-audio pairs, offering a heightened level of customization and specificity.

Classifier Guided Model. We train a separate classifier that can identify the video-audio using captions. The classifier training is a supervised learning using labeled video-caption pairs. The unconditional model is guided using the score of the classifier given input caption y. To this end, we employed the same U-Net blocks utilized in the main generative model. However, two key differences were introduced: 1) the output blocks of the U-Net were removed, and 2) a pooling layer and a fully connected layer were added to serve as a classification head. Initially, we focused on training a video classifier alone, but encountered discrepancies between the generated video and audio. Consequently, we devised a joint video and audio classifier. It constitutes a type of noisy classifier capable of accepting noised samples and the time step t as inputs. This element is of paramount importance, as the classifier plays a pivotal role in guiding each de-noising step of the diffusion backward process. Once we have a trained classifier, we leverage its guidance for the unconditional generative model. The unconditional generative model can be expressed as $P_\theta(v_{t-1}, a_{t-1}|(v_t, a_t))$, where v and a refer to video and audio, respectively. On the other hand, we define the joint video-audio classifier as $P_\phi(y|(v_{t-1}, a_{t-1}))$, with y denoting the input caption. Consequently, the classifier-guided generation can be expressed as:

$$\begin{aligned} &P_{\theta,\emptyset}(v_{t-1}, a_{t-1}|(v_t, a_t), y) = \\ &Z \cdot P_\theta\left(v_{t-1}, a_{t-1}|(v_t, a_t)\right) \cdot P_\phi(y|(v_{t-1}, a_{t-1})) \end{aligned} \tag{8}$$

Where Z is a normalizing constant, which is used to make any probability density function to have a total probability 1.

Conditional Diffusion Model with Caption Embedding. The Classifier-guided model, while effective, may incur significant time and computational costs when training on large datasets. As an alternative, we adopt the other approach in [30] by directly incorporating caption embeddings into the diffusion training process, eliminating the need for a separate classifier. Since classifier requires separate classifier that contains many parameters, which will increase the inference time. We finally selected to use conditional diffusion model. Specifically, we set $y = [y_t, y_c]$ as time step and caption embedding. To seamlessly integrate caption embeddings into each U-Net block, we employ the adaptive group normalization (AdaGN) layer [31]. By doing so, we establish the captioned conditional layer as follows:

$$AdaGN(h, y) = y_t GroupNorm(h) + y_c \tag{9}$$

Where h is the activations after the first convolutional layer of each block.

D. Speech Text Conditional Generation

For speech text conditional generation, our fundamental approach involves utilizing a pre-trained text-to-speech model to convert the given input text into speech audio. Subsequently, we harness this text-converted audio to guide the video generation process. we employ the YourTTS [21] model, a versatile multi-speaker and multi-language text-to-speech model. The incorporation of YourTTS enhances the flexibility and expressiveness of our model, empowering us to produce compelling video content guided by diverse textual inputs.

The guided process is the same as [32]. We define the text converted speech audio from pre-trained YourTTS as $\widehat{a}_0$. Then we can get $\widehat{a}_t$ using:

$$q(\hat{a}_t \mid \hat{a}_0) = \mathcal{N}(\hat{a}_t; \sqrt{1-\beta_t}\hat{a}_0, \beta_t I) \tag{10}$$

Once we have $\widehat{a}_t$, we replace a_t with $\widehat{a}_t$ in the backward processing. We then have $P_\theta(v_{t-1}, a_{t-1}|(v_t, \hat{a}_t), y)$. Finally, we updated the video $\tilde{v}_{t-1}$ using:

$$\tilde{v}_{t-1} = v_{t-1} - \mathrm{C} \cdot \nabla_{v_{t-1}} \|\hat{a}_{t-1} - a_{t-1}\|_2^2 \tag{11}$$

Where C is a constant for the conditional scale that can be set by the user.

E. Static Image Conditional Generation

Incorporating static images for conditional generation involves two distinct cases in our project: one with speech text input and the other without speech text input.

In the case with speech text input, we employ the text-converted speech audio to drive the static image using the pre-trained Wav2Lip [23] model. The Wav2Lip accepts image and audio as input, in most of the cases, the inputs are not synchronized. It can generate synchronized video using their synchronization expert model. It can accept both latent and audio/image native format as inputs. This allows us to synchronize the static image with the corresponding speech audio, producing a coherent and immersive video-audio combination. For the latter case, where speech text input is absent, we generate the audio using random distribution and subsequently use it to drive the static image. By leveraging this approach, we can still achieve dynamic and engaging video content, even in scenarios where specific speech text input is unavailable.

F. The Use of Pre-trained Models

To ensure user flexibility and enhance the quality of generated outcomes, we integrated two pre-trained models: YourTTS [21] and Wav2Lip [23].

YourTTS emerges as a powerful zero-shot text-to-speech audio model, boasting pre-training with multiple speakers and languages. It takes inputs of texts, language IDs, and speaker IDs, rendering it an essential component of our project. Given that our model was exclusively trained using English speech audio, we establish a fixed language ID set to English. Regarding speaker ID, in the absence of caption input, we randomly select a speaker index from the entire pool of speakers. However, when a caption is provided, the choice of female or male speaker index is determined accordingly. The resulting audio output is converted into a tensor format, serving as a guiding element for the subsequent video generation process. The significance lies in achieving text-to-speech video-audio generation without inflating the training parameters of the diffusion model. Indeed, the utilization of YourTTS streamlines our project and empowers us to achieve video-audio generation with enhanced flexibility. By incorporating the capabilities of YourTTS, we can readily adapt and generate speech audio from diverse texts, pre-defined language IDs, and pre-defined speaker IDs. The pre-defined language IDs and speaker IDs are limited to the YourTTS training data. This flexibility enhances the versatility of our system, enabling us to cater to a wide range of applications and scenarios. With YourTTS as an integral part of our project, we can effortlessly produce video-audio content with various linguistic nuances and speaker characteristics, significantly enriching the overall user experience.

Given that our project revolves around generating synchronized video-audio pairs for talking heads, our primary focus lies in achieving precise lip motion and audio synchronization. However, the subtlety of lip motions poses a challenge, making synchronization difficult. Additionally, discrepancies in lip-sync can be easily noticeable in videos compared with natural ambient sounds like ocean waves or wind, impacting the overall viewing experience. To address these concerns, we have integrated the pre-trained Wav2Lip model into our system. This model is specifically designed to enhance lip pose and audio synchronization. Wav2Lip leverages a pre-trained Lip-synchronization expert during training, with the objective of minimizing synchronization loss. We strategically incorporate this model after the final de-noising step, enabling us to significantly improve the synchronization of generated audio-video pairs. By leveraging the capabilities of Wav2Lip, we aim to produce talking head videos with remarkably accurate lip motion and audio coherence, delivering a seamless and immersive viewing experience.

V. Implementation Details

In this section, we provided an overview of several implementation details pertaining to our proposed model.

A. Diffusion Training

In the U-Net setting, we configured the number of ResNet [33] blocks for each U-Net block as 2, and the number of head channels was set to 64. The video and audio fps were chosen as 16 and 16,000, respectively. A fixed learning rate of 0.0001 was employed, and model saving was performed every 5,000 time steps during the training process. To schedule the diffusion noise, we utilized the linear method, the total diffusion steps T during inference is set as 1,000. To ensure meaningful outcomes, the model underwent training for a minimum of 50,000 time steps. The primary objective during training was to minimize the loss function between the predicted noise and the ground-truth noise. Furthermore, all video inputs were resized to 64×64. We define B as the batch size, F is the video frames, C is the channel number and A_d as input audio data points. Therefore, the input video tensor size is $[B, F, C, 64, 64]$ and audio tensor size is $[B, 1, A_d]$.

B. Classifier Training

During the training of the classifier, we established the initial learning rate as 0.001 and implemented learning rate decay to optimize its performance. To prevent overfitting and enhance generalization, we set the weight decay to 0.05. Given the classifier's role in handling noised input, we utilized the uniform method as the schedule sampler. We conducted a total of 300,000 iterations to ensure robust convergence and the acquisition of meaningful classification results.

C. Inference Setting

During the inference phase, we established the total de-noising steps as 1,000 and employed a linear noise schedule to effectively guide the de-noising process. To ensure the highest possible generation quality, we opted for the DDPM reverse sampling method.

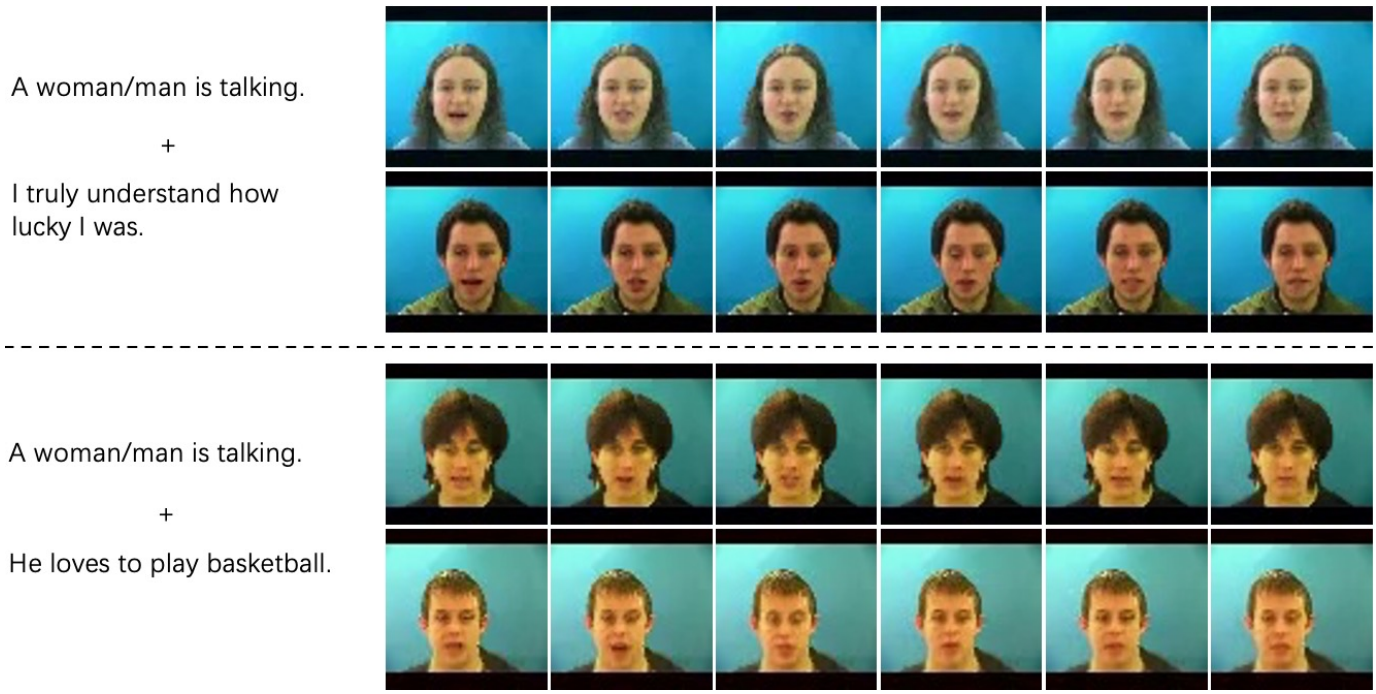

Fig. 4. Generated results of our method.

D. Dataset

We utilized a dataset introduced in [34], consisting of talking-head videos from 34 speakers, comprising 16 females and 18 males. To optimize training efficiency and cost, we selected a subset of 24 speakers, evenly distributed between 12 females and 12 males. Each speaker contributed 40 videos to the dataset. We further optimized training efficiency and cost by extracting one-second segments from each video, each featuring a single digit alongside one keyword. Consequently, the entire training dataset spans a duration of 16 minutes. To facilitate caption conditional generation, we meticulously labeled the videos with two distinct captions: "A man is talking" and "A woman is talking." The original resolution of each video is 360×288.

VI. Results

We present some sample frames generated by our model in Fig. 4. The model has the capability to accept two different video captions and any speech text to generate synchronized talking-head audio-video pairs. For additional generated talking videos, please refer to https://docs.google.com/presentation/d/1EqwP7UIr-brgw9A8L5Ww9sVd3dIMo_sH/edit?usp=drive_link&ouid=101411429973563845085&rtpof=true&sd=true.

VII. Limitation

The limitations of this work can be identified in the following aspects: 1) Achieving high-quality content generation with the diffusion de-noising technique requires numerous de-noising steps to reconstruct samples from pure Gaussian noise. However, the time-consuming nature of multiple de-noising steps may not be suitable for generating long videos efficiently, 2) While our model demonstrates the ability to generate natural talking head videos with dynamic head motion and eye blinking when provided with several static images representing continuous video frames as input (image batch condition), it tends to focus mainly on lip movements and exhibit limited head motion when using only one static image as input, 3) Due to GPU memory constraints, we had to resize the training video resolution to 64×64 pixels, which may be too low to satisfy the requirements of real-world content creation scenarios, 4) The model training process still relies on labeled video data for the caption conditional model. Presently, our methods accept only two kinds of captions, which may limit the diversity of generated content. These limitations highlight areas for future research and improvement in our proposed approach.

VIII. Conclusion

In conclusion, our work successfully realizes the task of text-to-joint talking-head video and audio generation, offering users a range of options for producing synchronized video-audio pairs. With the model's capability to accept any speech texts as input, it holds considerable promise for a wide array of everyday user applications. Moreover, the adoption of low-bitrate text inputs notably mitigates transmission costs during the generation process. This contribution underscores the practicality and versatility of our proposed approach in facilitating accessible and cost-effective multimedia content creation.

For **future work**, we can focus on expanding video caption categories, improving generation speed, fine-tuning model for unlabeled data guidance, and enabling talking head video generation from a single static image.

REFERENCES

[1] Ho, J., Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, pp.6840-6851.

[2] Li, B., Qi, X., Lukasiewicz, T. and Torr, P., 2019. Controllable text-to-image generation. Advances in neural information processing systems, 32.

[3] Qiao, T., Zhang, J., Xu, D. and Tao, D., 2019. Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1505-1514).

[4] Zhang, Z. and Chang, M.C., 2023, August. Two-stage dual augmentation with clip for improved text-to-sketch synthesis. In 2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 1-6). IEEE.

[5] Bhise, N., Zhang, Z. and Bui, T.D., 2020. Improving text to image generation using mode-seeking function. arXiv preprint arXiv:2008.08976.

[6] Li, Y., Min, M., Shen, D., Carlson, D. and Carin, L., 2018, April. Video generation from text. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[7] Hong, W., Ding, M., Zheng, W., Liu, X. and Tang, J., 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868.

[8] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. and Ho, J., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35, pp.36479-36494.

[9] Tumanyan, N., Geyer, M., Bagon, S. and Dekel, T., 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1921-1930).

[10] Zhang, C., Zhang, C., Zhang, M. and Kweon, I.S., 2023. Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909.

[11] Zhang, L., Rao, A. and Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3836-3847).

[12] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J. and Norouzi, M., 2022. Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence, 45(4), pp.4713-4726.

[13] Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q. and Guo, B., 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10219-10228).

[14] Alexanderson, S., Nagy, R., Beskow, J. and Henter, G.E., 2023. Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG), 42(4), pp.1-20.

[15] Amit, T., Shaharbany, T., Nachmani, E. and Wolf, L., 2021. Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390.

[16] Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P. and Cattin, P.C., 2022, December. Diffusion models for implicit image segmentation ensembles. In International Conference on Medical Imaging with Deep Learning (pp. 1336-1348). PMLR.

[17] Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B. and Zhu, L., 2023, October. Diffmic: Dual-guidance diffusion network for medical image classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 95-105). Cham: Springer Nature Switzerland.

[18] Mukhopadhyay, S., Gwilliam, M., Agarwal, V., Padmanabhan, N., Swaminathan, A., Hegde, S., Zhou, T. and Shrivastava, A., 2023. Diffusion models beat gans on image classification. arXiv preprint arXiv:2307.08702.

[19] Bandara, W.G.C., Nair, N.G. and Patel, V.M., 2022. Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models. arXiv preprint arXiv:2206.11892.

[20] Li, Z., Huang, Y., Zhu, M., Zhang, J., Chang, J. and Liu, H., 2024. Feature manipulation for ddpm based change detection. arXiv preprint arXiv:2403.15943.

[21] Casanova, E., Weber, J., Shulby, C.D., Junior, A.C., Gölge, E. and Ponti, M.A., 2022, June. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In International Conference on Machine Learning (pp. 2709-2720). PMLR.

[22] Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X. and Zhao, Z., 2023, July. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In International Conference on Machine Learning (pp. 13916-13932). PMLR.

[23] Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y. and Adi, Y., 2022. Audiogen: Textually guided audio generation. arXiv preprint arXiv:2209.15352.

[24] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V., 2020, October. 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).

[25] Wang, S., Li, L., Ding, Y., Fan, C. and Yu, X., 2021. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. arXiv preprint arXiv:2107.09293.

[26] Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y. and Xu, C., 2020, August. Talking-head generation with rhythmic head motion. In European Conference on Computer Vision (pp. 35-51). Cham: Springer International Publishing.

[27] Norris, J.R., 1998. Markov chains (No. 2). Cambridge university press.

[28] Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing.

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[30] Dhariwal, P. and Nichol, A., 2021. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, pp.8780-8794.

[31] Nichol, A.Q. and Dhariwal, P., 2021, July. Improved denoising diffusion probabilistic models. In International conference on machine learning (pp. 8162-8171). PMLR.

[32] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M. and Fleet, D.J., 2022. Video diffusion models. Advances in Neural Information Processing Systems, 35, pp.8633-8646.

[33] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[34] Cooke, M., Barker, J., Cunningham, S. and Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5), pp.2421-2424.

[35] Ramachandran, P., Zoph, B. and Le, Q.V., 2017. Searching for activation functions. arXiv preprint arXiv:1710.05941.