論文 / 著書情報
Article / Book Information

題目(和文)	音声と映像の関連性を活用した現実的な音声駆動型話者顔合成とその応用
Title(English)	Realistic Speech-Driven Talking Face Synthesis via Audio-Visual Association Exploitation and its Applications
著者(和文)	SunYasheng
Author(English)	Yasheng Sun
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第25号, 授与年月日:2024年12月31日, 学位の種別:課程博士, 審査員:小池 英樹,篠田 浩一,岡崎 直観,齋藤 豪,井上 中順
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第25号, Conferred date:2024/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Doctoral Dissertation

Realistic Speech-Driven Talking Face Synthesis via Audio-Visual Association Exploitation and its Applications

Yasheng Sun

November 21, 2024

Computer Science Course
Department of Computer Science
School of Computing
Tokyo Institute of Technology

A Doctoral Dissertation
submitted to School of Computing,
Tokyo Institute of Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Yasheng Sun

Thesis Committee:

Professor Hideki Koike (Supervisor)
Professor Naoaki Okazaki
Professor Koichi Shinoda
Associate Professor Suguru Saito
Associate Professor Nakamasa Inoue

Realistic Speech-Driven Talking Face Synthesis via Audio-Visual Association Exploitation and its Applications *

Yasheng Sun

Abstract

Talking face synthesis holds significant importance across various entertainment applications, including digital human animation, visual dubbing in films, and rapid production of short videos. To promote realistic video generation, we propose to fully exploit the implicit information of human speech in talking face synthesis system. In contrast to previous works that borrow external images, videos or labels as reference source to explicitly provide speaker characteristics like facial appearance or talking status, our proposed framework attempt to *directly infer these underlying features from human voice, thereby increasing realness and naturalness of the synthesized videos.*

However, this task is particularly challenging due in two aspects. 1) These information within human voice is incomplete or even ambiguous. Such uncertainty brings difficulty to development of talking face system because capturing the required information underlying human speech is non-trivial. 2) Successfully driving a talking face requires delicately integration of interleaved information such as speech content or speaker characteristics from human voice. This further increases the complexity of this task. To address the above challenge, we propose to *leverage the explicit visual cues as a bridge to* promote audio-visual association learning, thus facilitating extensive exploitation of human speech. Specifically, our framework is divided into three components. To ensure that the generation progress is well delegated to each specific semantic space and prevent undesirable interference among them, we first carefully derive a strategy to disentangle the

*Doctoral Dissertation, School of Computing
Tokyo Institute of Technology, November 21, 2024.

latent space of our generator into speech content space and content-irrelevant space. Then, we leverage explicit visual cues to bootstrap the implicit vocal representations via maximizing their mutual information. Finally, to further narrow down the representation disparity of audio-visual modality, we instruct the system adapt to the audio representation, thereby allowing smooth transition from usage of explicit visual cues to implicit vocal features.

We develop two applications to validate the effectiveness of our proposed framework on both 2D and 3D talking face synthesis, respectively. In our first application, we attempt to directly infer the appearance of a person's face by listening to the speaker's utterance. To animate it, we also need to leverage identity-irrelevant information such as speech content. Therefore, we first exploit visual cues to disentangle identity space that models bio-metric appearance, and identity-irrelevant space in a style-based generative framework. Then the vocal feature is introduced to synchronize visual representation from two perspectives. To automatically balance those information from two spaces, the generator is delicately fine-tuned following curriculum learning paradigm. In second application, we target to incorporate the inherent talking status information conveyed by human speech. A pretrained wav2vec model is utilized to identify the speech content space and the content-irrelevant space is complementarily learned. To facilitate utilization of human voice, we leverage visual descriptions of talking video as bridge. Specifically, we leverage these visual instructions as input to guide 3D talking face generation within the identified speech-irrelevant space. Finally, human speech is contrastively aligned with and predict visual instructions for effective talking status representation.

Extensive experiments show that our approach encourages better speech-identity correlation learning while generating vivid faces whose identities are consistent with given speech samples. And the same model enables these inferred faces to talk driven by the audios. For the second application, we demonstrate that our approach produces vivid talking faces with expressive facial movements and consistent emotional status. To our best knowledge, this is the first speech-driven talking face study that incorporating implicit vocal information bridging by explicit visual cues.

Keywords:

Video Generation, Audio-Visual Correlation, Talking Face Synthesis, Contrastive Learning, Lip Synchronization, Large Language Model, Realistic Generation

Acknowledgements

First and foremost, I extend my deepest gratitude to my advisor, Prof. Hideki Koike, for his unwavering support and guidance throughout my doctoral journey. Despite my background in mechanical engineering, Prof. Koike generously provided me with the opportunity to pursue my research interests in computer science. His kindness and open-mindedness created a conducive research atmosphere and provided me with a platform to explore my areas of interest, constantly motivating me to strive for excellence. In the academic realm, his unique perspective and meticulous attitude to high-quality presentation have inspired me to strive for clearer expression and articulate my opinions more effectively.

I would also like to convey my deepest gratitude to my dissertation committee members: Prof. Naoaki Okazak, Prof. Koichi Shinoda, Prof. Suguru Saito, and Prof. Nakamasa Inoue. Their careful examination of my dissertation, insightful comments, and constant encouragement have been invaluable and their questions during the interim presentation and pre-examination have encouraged me to widen my research from varied perspectives.

I am profoundly thankful to all my lab members and friends for their camaraderie and support throughout my graduate studies. Special thanks to Shio Miyafuji and Setsuko Mizoguchi for their invaluable assistance and accompany. The vibrant and collaborative atmosphere fostered by my lab mates, including Jefferson Pardomuan, Keishiro Uragaki, Liao Chen Chieh, Daichi Saito, Jana Hoffard, Luna Takagi, Ruofan Liu, Eruku Iida, Yuka Tashiro, Takashi Matsumoto, Yuki Sato, Zhihao Yu, Hidetaka Katsuyama, Arisa Kohtani, Toshiki Omi, Kyohei Hayakawa, Yusuke Kojima, Takuya Takahashi, Toshihiro Hirano, Yuka Tanaka, Takuto Nakamura, Dong-Hyun Hwang, Luna Takagi, Hui-Shyong Yeo, Christopher Mitcheltree, Mikihito Matsuura, Kohei Aso, Xuan Zhang, Toshiki Sato, Nobuhiro Takahashi, Haruki Kikuchi, Yusuke Miura, greatly enriched my academic experience.

I am immensely grateful for the academic guidance provided by Prof. Ziwei Liu

from Nanyang Technological University. Additionally, I owe a debt of gratitude to my friends who selflessly supported me in both my life and research endeavors. Haoxiang Shi from Waseda University, Ao Liu, An Wang, and Zhishen Yang from Okazki Lab of Tokyo Institute of Technology and Bohan Li from Shanghai Jiao-Tong University have been invaluable sources of encouragement and inspiration.

Furthermore, I express my sincere appreciation to the teams at Sensetime Research, Netease Fuxi Lab, Baidu VIS, and Microsoft Research Asia for their wholehearted support and insightful discussions about my research work. Special thanks to Hang Zhou, Kaisiyuan Wang, Wenqing Chu, Zhiliang Xu, Jiangke Lin, Wenming Qian, Yi Yuan, Yifan Yang, Houwen Peng, Han Hu, and others for their invaluable contributions.

Lastly, I extend my heartfelt gratitude to my family for their unwavering love and support throughout my life. Their encouragement and sacrifices have been instrumental in shaping my academic journey, and I am deeply grateful for their presence in my life.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

1.1.1 Talking Face Generation

The widespread adoption of virtual humans across multiple sectors such as personal assistance, intelligent customer service, and online education owes much to the rapid advancements in artificial intelligence [178]. These anthropomorphic digital beings have the ability to swiftly engage with individuals, thereby augmenting the user experience within human-computer interactions. A virtual character, designed to emulate human behavior, is anticipated to exhibit appropriate body movements, hand gestures, and communication skills. Notably, the vividness of facial animations plays a pivotal role in shaping the human experience, particularly during verbal exchanges, which serve as a key mode of interaction. Therefore, the talking face generation has become an active research field where the vocal consistency of lip shapes and facial attributes, such as facial expressions and eye movements are taken into account. This task targets to synthesize lip-synced talking face video by accepting an human speech as input as well as other references such as language instruction, emotional labels, facial image or another video providing extra control information. Strictly speaking, talking face generation with head movement is also called talking head generation [85]. Usually, besides the vivid facial details, there are also vivid rhythmic head movements in the talking head generation results. In this work, we call both of them talking face synthesis without differentiate them in purpose for convenience.

Earlier works in talking face synthesis attempt to learn a mapping from vocal features to visual features using traditional methods [161, 146] such as DBN (Dynamic Beyesian Network) or HMM (Hidden Markov Model). The relationship between human speech and video units is modeled by basic computer graphic approach [15]. A strategy is derived to selectively search the most relevant visual

features and the transition between clips is achieved by naive interpolation. But this kind of simplification is prone to cause unnatural lip movements. Later work [161] improves the smoothness and synchronization by designing the audio-visual articulatory model based on the Dynamic Bayesian Network (DBN) to directly simulate the movement of the articulators including lips, tongue, and teeth. To cultivate photo-realistic talking applications, a statistical HMM [146] is trained to automatically select the mouth sequences from a pre-recorded database according to audio sequences and fuse it to background. Since relying on the pre-recorded database, their application is limited to this specific person.

Nowadays, data-driven approaches, particularly those that are neural network-based, have taken center stage in this domain [3, 21, 40, 122, 139, 189, 131]. Various deep learning based talking face applications have been developed. Those works that target for specific talking face generation, similar to [146], also collects numerous video clips of one person and leverage a neural network such as LSTM [40] or RNN [131] to map audio features to his lip animation of mouth regions. These slew of studies, relying on implicitly learning construction of facial model [131, 67], could accomplish realistic synthesis with superior quantitative and qualitative performance. But it requires manual collection of video clips from a specific person and faces challenges to adapt to other target identities, limiting its applications.

Another slew of works [75, 110] attempt to extend it to arbitrary talking face generation. Some of these works [134, 20, 75, 110, 37] focus on synthesizing the lip movements around mouth region where they mask out the mouth area and inpaint the region according to vocal rhythm. The generated mouth area is pasted back to the original video seamlessly for realistic syntheis. It has been proved that by learning joint audio and visual embedding of the lower part face, LipGAN [75] can generate a mouth consistent with provided upper face pose and audio. His following work, Wav2Lip [110], further improves the lip-synchronization performance by introducing a pretrained lip-sync discriminator. With the powerful capability of Generative Adversarial Network (GAN), these approaches [20, 37] further improve the image quality and achieve more realistic synthesis. However, only taking into consideration of correlation of speech audio and mouth region, they are unable to synthesize the whole face, thereby eye movements or emotional information are neglected. To cover facial regions, other works [155, 185, 123, 62] target to leverage the audio features to predict visual features of whole face. In-

Method	Year	Identity Source	Emotional Source	Representation
Mead [144]	2020	Ref Image	One-hot Label	2D
GC-AVT [80]	2022	Ref Image	Ref Video	2D
EAMM [65]	2022	Ref Image	Ref Video	2D
Sinha *et al.* [121]	2022	Ref Image	One-hot Label	2D
TalkClip [90]	2023	Ref Image	Language	2D
MeshTalk [113]	2021	Ref Mesh	Ref Mesh	3D
CodeTalker [163]	2023	Ref Mesh	N/A	3D
EMOTE [34]	2023	Ref Mesh	One-hot Label	3D
Speech2Talking-Face	2021	**Human Voice**	N/A	2D
AVI-Talking	2024	Ref Mesh	**Human Voice**	3D

Table 1.1: Approaches of talking face via deep learning.

termediate representations [21, 109] such as facial landmarks and facial models are introduced to provide a overall guidance in the generation process. These approaches leverage RNN [94] or LSTM [55] network to learn a mapping from audio features to the structural representation, implicitly modeling the whole face dynamics. But construction of these structural representation is prone to introduce error accumulation. Some researches [181] directly learn a disentangled audio-visual latent representation and thus able to synthesize more realistic mouth movement results. However, without explicit modeling of audio information association with facial details, they could only provide a global dynamics, neglecting fine-grained degree of movements such as eye movement, muscle line changes, eyebrow changing or cheek raising.

1.1.2 Realistic Talking Face Synthesis

To foster more realistic synthesis, later researchers pay more attention to facial details other than mouth region. As crucial component of visual animation, blinks and eyebrow movements has become an active research topic. Earlier works [140, 120, 172] leverage GAN-based architecture to unconditionally learn an blink action driven by random noise. Conditioned on random noise, the synthesized eye movements cannot be freely controlled and often suffer from simple

blink motion with naive motion pattern. For better controllablity, Hao et al. [51] devised a method enabling controllable blinking actions in talking face generation, employing a blink conversion network and joint training to enhance conversion effectiveness. Their following work [86] introduced a feature-driven architecture facilitating direct control over blink actions in high frame rate videos, ensuring blink features govern blink actions independently of facial and identity features. Some other works leverage one-hot label [175] or a reference video clips [80] to manipulate the movements around eyes. Despite accomplished controllability, it requires external manual interference with assigned reference video or labels.

Therefore, a more natural way is to directly leveraging relevant information from audio. Liu and Wang [87] proposed a two-stage method for generating detailed talking faces by mapping speech audio features to action units, achieving audio-driven conversion and producing more realistic videos with richer facial details. Lacking of diverse facial action units in their utilized dataset, their produced results still fail to capture fine-grained semantic details. As MEAD (Multi-view Emotional Audio-visual Dataset) [144], a large-scale dataset for emotional talking face generation, is collected, later works [144, 121, 38, 65] are capable of generating emotion-aware talking faces with explicit control. But these works either rely on categorical one-hot label [144, 121, 38] or external reference [65]. In Table. 1.1, we briefly summarize those approaches that attempt to synthesize talking face with realistic details. Most of them propose to leverage external features to control the generation progress. A natural way of directly utilizing information from speech is worth exploring.

1.2 Research Summary

1.2.1 Inspirations and Challenges of Realistic Synthesis

Driving a face to talk with either a clip of audio [21, 189, 181, 110, 122, 186, 64, 22, 183] or video [155, 170, 14, 72] has long been the topic of interest for both the fields of computer graphics and artificial intelligence. Earlier methods focus more on modeling a specific person [131, 72]. Recent studies propose to achieve arbitrary subject or one-shot talking face generation driven by audios or videos. A number of studies borrow intermediate representations such as landmarks [170, 21] into this task. But this is not convenient in our settings, as voice-reconstructed faces

have no fixed landmarks. Certain landmark-free methods [181, 110, 183] rely on skip-connections thus require an image as network input. Specifically, [14] disentangles facial identities and poses in a style-based framework. However, they require a meta-learning stage on ground truth videos.

It has been proven in neuroscience and psychology [66, 93] that detailed facial appearances can somehow be inferred from human voices. The same genetic, physical or environmental factors may biologically affect both voice and face [152]. Based on this knowledge, recent studies in artificial intelligence have sought to associate facial appearances with the human speech in a data-driven manner. They propose to retrieve faces with speech [98, 71, 97] and even directly generate face images from speech [100, 153, 24], to fulfill human's visual imagination when hearing a voice. However, previous studies focus more on depicting a still face image. They normally match audio features with visual ones [100, 153, 24], and then leverage a pretrained face normalization or generation network for reconstruction. But facial appearance is not the only thing that we can imagine when listening to a clip of speech. We shall also picture the mouth movement which is strongly correlated with the content of the speech. Meanwhile, researchers have succeeded in driving a known facial image to speak conditioning on audios [181, 21], but the reference identity has to be given. The ability to *directly generate the appearance and movements of a face could be useful in multiple scenarios*, such as animating virtual humans solely from a piece of speech. This is a non-trivial task. On one hand, there is no fixed matching between voice and faces. It is not possible to reconstruct the exact face given a clip of speech. On the other hand, identity and lip motion information are entangled within both faces and audios. It is difficult to build supervision for both information individually at the same time.

Another often overlooked aspect of human speech is the implicit speaker status it conveys. Humans possess the innate ability to infer the perceived speaking status of an individual based on their voice. To synthesize expressive speech-driven 3D talking face, previous work either 1) model the correlation between dynamic head poses and audio rhythm [22, 129] or 2) borrow an external representation [91, 90, 108] such as emotion labels or video clips as style reference during generation. However, the head dynamics hold limited expressive ability thus only yield coarse alignment, neglecting the emotional nuances present in the audio content. The latter studies require manual style source selection by users, leading to unnatural applications. In the paper, we explore a more natural

scenario, targeting to *directly leverage the underlying style information conveyed by human speech for generating an expressive talking face that aligns with the speaking status.*

Synthesizing diverse and plausible facial details based on speech while maintaining accurate lip synchronization is a highly challenging task. This challenge stems from the inherent ill-posed nature of the problem, characterized by 1) one-to-many relationship between audio inputs and potential facial movements consistent with the spoken content. Some efforts [22, 169, 92] have introduced diffusion mechanisms to tackle diverse generation. However, direct diffusion from audio to facial motion requires bridging a huge modality gap while the information within speech and facial movements are often weakly correlated. With heavy learning burden and limited model capability, such practice is prone to capture only coarse alignment with audio cues, neglecting emotional nuances of the speaker. 2) The intertwining of the speaker's talking style and lip movements further complicates the synthesis process. Prior work [91] aimed to address this entanglement by controlling specific coefficients of a parametric model. However, such practice relies on a disentangled parametric model, which is not always accessible or precise.

In summary, few studies have delved into leveraging the fine-grained vocal characteristics of speakers, including personal identity and speaking status, within the realm of talking face synthesis. This serves as our motivation to develop a suitable strategy aimed at fully uncovering these nuanced vocal attributes and harnessing them to enhance the realism of synthesis.

1.2.2 Research Objective

In this study, we delve deeper into the intricate correlations between speech and facial expressions, striving to extract as much detailed information as possible from a speech clip. Our goal is to advance the realism of talking face synthesis by enhancing the consistency between synthesized faces and the input human speech. Specifically, our approach aims to imaginatively capture the speaker's unique vocal characteristics and reflect them in the generated speaker identity. Additionally, we seek to explore the implicit speaking status embedded within human voice and ensure that the synthesis of facial details aligns convincingly with this context.

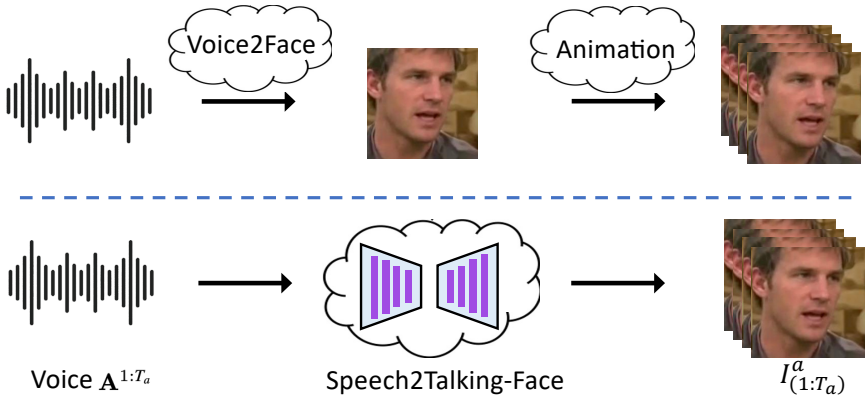

Figure 1.1: Rather than adhering to a two-stage animation paradigm, we aim to directly visualize a talking face from human voice.

1.2.3 Proposed Solutions

This study explores effective approaches to enhance the realism of talking face synthesis by addressing the aforementioned issues, namely: (1) The autonomous adaptation of the person's identity in the talking face to match the input human speech, and (2) ensuring consistent facial details that correspond to the speaking status conveyed by human speech. The key objective is to *effectively extract these pieces of information from the voice and seamlessly integrate them into the synthesized videos*. To establish fine-grained audio-visual associations, we meticulously design a training strategy from both recognition and generative perspectives. Specifically, we employ contrastive learning to fully exploit the distinctive features in human voice by maximizing mutual information. As for the generator, we propose initially identifying a relevant space through feature disentanglement, followed by learning to utilize these extracted pieces of information within this identified feature space.

For the incorporation of speaker identity, we propose **Speech2Talking-Face**, a generative framework that is not only capable of inferring appearances but also driving the faces to talk with speech. As depicted in Fig. 1.1, our study diverges from the two-stage animation paradigm. Rather than sequentially generating a plausible face from human voice and then animating it with audio using a speech-driven network, we aim to accomplish both tasks within a unified framework, ensuring comprehensive integration of speech information. The key insight is

to *synchronize speech with visual representations in two branches, then utilize a style-based generator for multi-modal information balancing.* Specifically, two latent spaces, namely the identity space and the identity-irrelevant space, are defined through the reconstruction training of visual features as [14, 183]. Proven to be effective for learning coherent audio-visual representations and cross audio-visual generation [97, 28, 181, 182, 110, 47], the contrastive loss is leveraged for our modality synchronization in the two spaces. Detailedly, due to the fact that identity-irrelevant space contains speech-irrelevant information such as head poses, we map speech features to a *content* subspace in the identity-irrelevant space by synchronizing speech with the differences of visual feature. In this way, the initial pose of our generated face can be determined by any facial image. Furthermore, inspired by discoveries in face recognition [35], class centroids of identities are leveraged for more compact identity representation learning.

Finally, the features from the two spaces are integrated in the style-based generator. Visual-to-audio curriculum learning is adopted to support the generator in balancing information from two latent spaces and two modalities. After training, our model is capable of leveraging one clip of speech to generate a dynamic face whose appearance is controlled by the *identity* space, with accurate lip motion controlled by the *content* subspace. Compared with previous methods, our generated results are of higher quality and more attractive given our particular ability to reflect speech content information.

For the incorporation of speaker talking status, we present **AVI-Talking**, an **A**udio-**V**isual **I**nstruction System for expressive **Talking** face generation. Our key insight is to *bridge the huge audio-visual modality gap with an intermediate visual instruction representation.* As shown in Figure. 1.2, in contrast to previous approaches that directly learn facial movements from audio, our framework decomposes the audio-to-video generation into two stages, each with a distinctive objective, thus significantly mitigating the optimization complexities. Specifically, while speaker voice entails complex information, language instruction typically conveys clearer meaning. This inherent clarity enhances the performance of the synthesis network, leading to superior results. To facilitate this, we integrate Large Language Models, leveraging their contextual reasoning capabilities to comprehend human speech and simulate plausible speaker states. By separating the generation and understanding functions, we ensure specialized expertise is responsible for each task. Furthermore, by presenting visual instruction as an

intermediate output, our system not only enhances model interpretability but also grants users the flexibility to specify desired instructions or modifications. This feature enriches user interaction and greater customization.

As depicted in Fig. 1.2, the first stage aims for comprehending the speaker talking state and imaginatively generate plausible facial expression details for subsequent instruction, necessitating robust contextual reasoning and hallucination capability. Inspired by the impressive multi-modal understanding and generation abilities demonstrated by recent large language models (LLMs) [159, 5], we propose integrating LLMs as an agent [145] to guide the talking face synthesis process. The key aspect lies in *formulating a soft prompting strategy to harness the prior contextual knowledge underlying LLMs* for speaker talking state comprehension. To achieve this, we initially employ a Q-Former to contrastively align speech features with visual instructions. Building upon the aligned audio features, we fine-tune a small number of parameters in the input projection layers for domain adaptation. Such practice not only facilitates efficient tuning but also promotes the utilization of language priors.

In the second stage, with the obtained visual instructions, our objective is to develop a speech-to-talking face network capable of synthesizing facial details that adhere to the provided instructions while preserving accurate lip movements. To address the inherent entanglement between lip movements and the speaker's talking style, we propose *deriving a compressed latent space that distinctly identifies features related to speech content and those irrelevant to content.* By integrating both types of latent features, we can reconstruct expressive facial movements through a talking generator, thereby bypassing issues associated with inaccurate or inaccessible disentangled parametric spaces [90]. In order to leverage this devised talking prior for instruction-following generation, it is crucial to align visual instructions within the *content irrelevant* space. To facilitate joint representation learning, we introduce a contrastive instruction-style alignment and diffusion strategy. Specifically, we initially align the visual instruction contrastively to the shared content irrelevant space, upon which a diffusion prior network is employed to further refine this joint representation towards the distribution of the pre-trained talking prior.

In summary, this study presents a unified framework aimed at fully leveraging the implicit information within human speech by enhancing fine-grained cross-modal associations. This framework facilitates two novel applications and

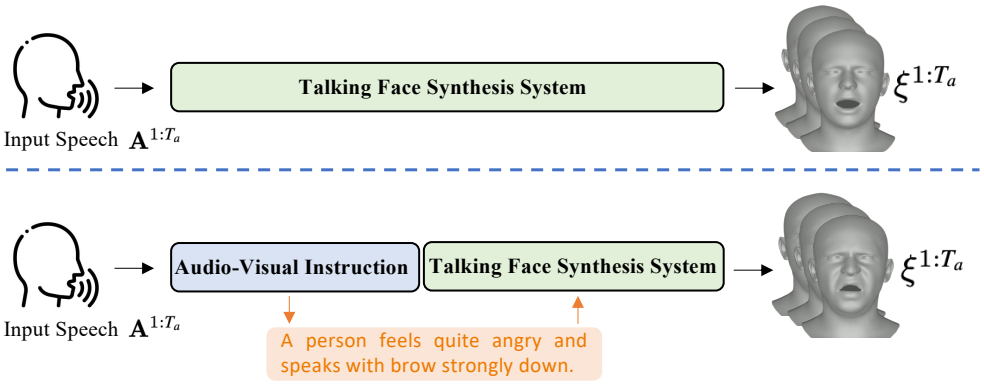

Figure 1.2: Unlike previous works, we introduce an audio-visual instruction module to instruct the talking face synthesis process.

showcases its effectiveness in both 2D and 3D talking face settings.

1.2.4 Contributions

The contributions of this study are as follows.

1. We propose to extract as much speech-facial associated information as possible with two branches of audio-visual synchronization. We strengthen the synchronized identity space with the concept of class centroids, and learning a content subspace within the identity-irrelevant space.

2. We successfully achieve speech-inferred facial appearance reconstruction and speech-driven talking face generation within a unified model called Speech2Talking-Face. To the best of our knowledge, this problem has not been addressed before.

3. We propose an innovative audio-visual instruction system, AVI-Talking, that decomposes expressive talking face generation into two stages: audio-visual instruction generation and facial movement synthesis. Experimental results validate the capability of AVI-Talking in generating vivid 3D talking faces with expressive facial details and a consistent emotional status.

4. To interpret the speaker's talking status, we introduce Large Language Models (LLMs) as agents for audio-visual instruction. They generate plausible speaker talking status based on the human speech.

5. For precise instruction-following synthesis, we introduce a language-guided talking prior network with disentangled speech content and content-irrelevant space. Additionally, we design a diffusion network to fully exploit the motion prior.

Through these contributions, this study advances the field of talking face synthesis by paving the way for more realistic outcomes. We provide a thorough understanding of the complete utilization of human speech information and its seamless integration into the synthesis process of a generative model. The effectiveness of our approach is demonstrated through the presentation of two practical applications, affirming the efficacy of our proposed strategy.

1.3 Thesis Outline

In the rest of this thesis, Chapter 2 describes the related work for realistic talking face synthesis. This includes the prior works on 2D talking face synthesis and speech-driven 3D talking head synthesis. The related work is described in two veins, datasets, and approaches. The datasets part, in particular, provides the details of the four datasets used in this study. In addition, Chapter 2 introduces the preliminary knowledge relevant to this study, including the definition of 3D parametric facial models, the StyleGAN architecture, the basic theory of diffusion models, and the pretrained models which serve as the feature extraction models in this study.

Chapter 3 provides a high-level overview of philosophy behind this realistic talking face system, including research motivation, design choice of our framework and in-depth analysis of common points behind our promoted applications.

Chapter 4 introduces the first application, Speech2TalkingFace, to improve the speaker identity consistency between visual synthesis and human speech input. This approach is focused on the task setting of synthesizing a talking face solely from a clip of audio, and aims to generate both lip-synchronized and speaker identity consistent videos.

Chapter 5 introduces the second proposed application, AVI-Talking, to improve the emotional consistency between visual synthesis and human speech input. This approach introduces the texts to reinforce the feature extraction of speech emotion

and leverages Large Language Models (LLMs) to instruct the 3D talking face synthesis process.

Chapter 6 concludes this thesis by summarizing the proposed methods. It also discusses the limitations of this work and the future directions.

2 Related Work and Preliminary

In this chapter, we provide an overview of the related work in talking face synthesis. This includes a thorough examination of general talking face generation techniques and research focused on enhancing the realism of synthesis methods. We also present a detailed description of commonly used datasets in the field. Additionally, we incorporate a range of studies that have influenced and served as foundational elements for our research, covering face-voice association learning, emotion recognition, and the use of large language models (LLMs) in cross-modality applications. Furthermore, this chapter offers an introduction to the essential preliminary knowledge necessary for a comprehensive understanding of our research.

2.1 Related Work on Talking Face Synthesis

2.1.1 Datasets

VoxCeleb2 [30]. VoxCeleb2 features 6,112 celebrities, spanning over 1 million utterances. Of these, 5,994 speakers are included in the training set, while the remaining 118 are allocated to the test set. Most videos are sourced from YouTube, resulting in a wide variance in video quality. Some videos exhibit extreme characteristics, such as significant head pose movements, low-light environments, and varying degrees of blurriness. Importantly, none of the identities in the test set have been encountered during the training phase. Figure 2.1 provides visual examples of speaking faces, shown in the left segment.

Lip Reading in the Wild (LRW) [29]. This dataset is specifically designed for lip reading, containing over 1,000 utterances of 500 distinct words in each 1.16-second video clip. Each video comprises 29 frames at a frame rate of 25Hz, with the target word positioned in the middle of the clip. Unlike VoxCeleb2, the

Figure 2.1: Samples from VoxCeleb2 dataset and LRW dataset [30, 29].

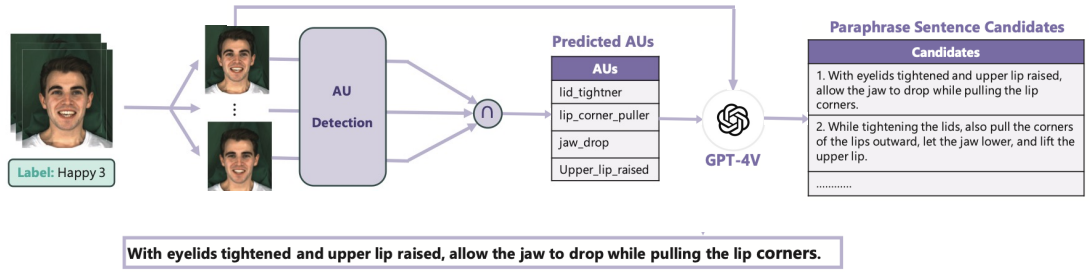

Figure 2.2: Description of InstructAvatar [149] dataset.

videos in this dataset are mostly clean, high-quality footage sourced from BBC news, featuring near-frontal faces. Consequently, a significant portion of both the utterances and identities in the test set are also encountered during training. Notably, the facial region cropping in this dataset includes more of the hair region. Visual samples of these facial images are provided on the right side of Figure 2.1.

InstructAvatar [149]. InstructAvatar dataset is devised by augmenting the MEAD [144] dataset. To extend the emotional labels, the Facial Action Units (AUs) are first detected to describe muscle movements as depicted in Fig. 2.2. Large language models, GPT-4V, are capitalized to connect these key words and paraphrase them with diverse natural language descriptions. Meanwhile, the extracted visual frame is also fed into GPT-4V to refine the obtained descriptions for consistent expressions.

MeadText [90]. MeadText builds upon the Mead dataset, as introduced in

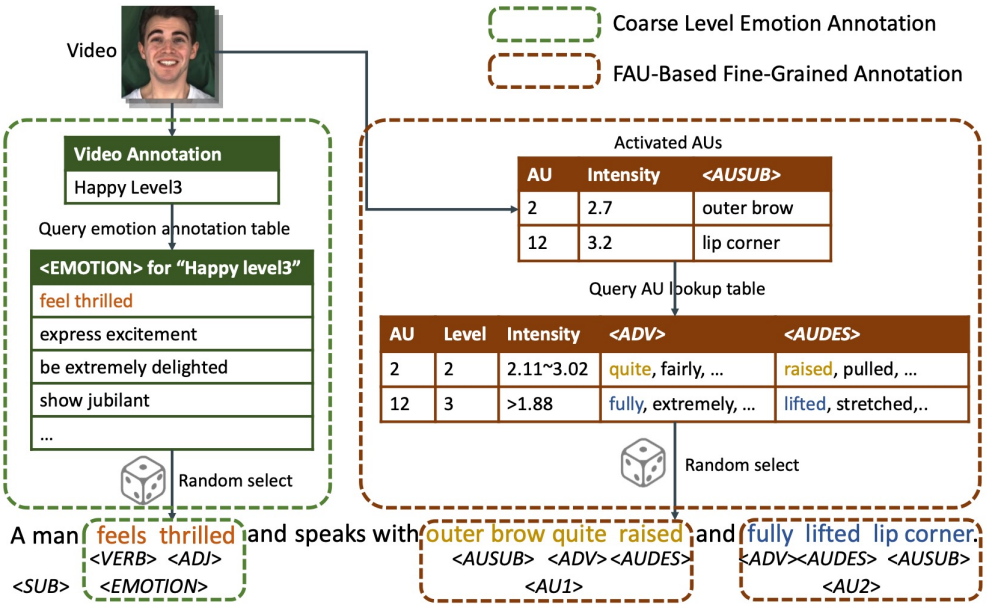

Figure 2.3: Description of MeadText [90] dataset.

Wang et al.'s work [144], which features speaking videos encompassing eight distinct emotional states. The annotation process for MeadText, illustrated in Figure 2.3, involves both coarse and fine-grained levels of emotion categorization. Specifically, annotators classify emotions into various levels, and a corresponding phrase associated with the labeled emotion is randomly selected for coarse-level annotation. For fine-grained annotation, off-the-shelf Facial Action Units networks (FAUs) are used to detect activated AUs. A lookup table is then employed to associate a phrase describing the detailed action, as depicted on the right side of the figure.

RAVEDESS [89]. The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a resource designed for emotion recognition research. It features recordings from 24 professional actors (12 female, 12 male) producing over 1,440 utterances in a neutral North American accent. The database encompasses eight speech emotions: calm, happy, sad, angry, fearful, surprised, disgusted, and neutral, with each emotion expressed at two levels of intensity: normal and strong. RAVDESS includes three modality formats: audio-only, audio-video, and video-only.

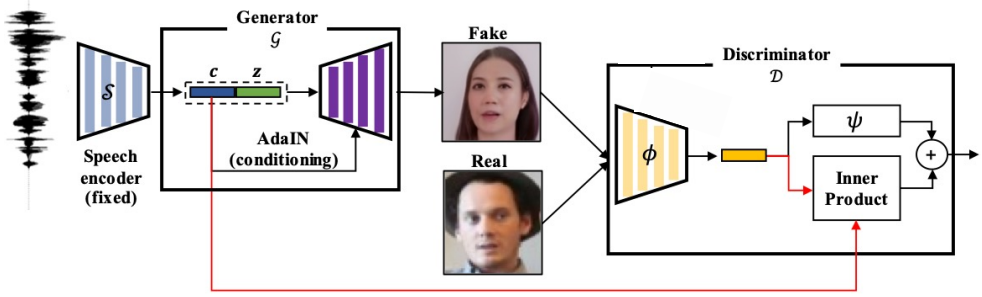

Figure 2.4: Vocal features pertinent to facial characteristics are identified in generative process [24].

2.1.2 Face-Voice Association Learning and Reconstruction

Studies on face-voice association learning [98, 97, 71, 57] aim to reveal the relationship between facial appearance and voice through cross-modal matching. Nagrani et al. [98] introduce both binary and multi-way cross-modal face and voice matching tasks via classification. In their subsequent work [97], they propose learning common cross-modal embeddings for person identity using a self-supervised framework with contrastive loss. Metric learning [71] and multi-task learning [152] are employed to develop a shared representation for different modalities. In our work, we also utilize a similar contrastive learning approach for our identity space embedding, demonstrating that our synchronized speech-identity representation benefits face-voice association learning.

Reconstructing faces from human voices has recently garnered significant attention in artificial intelligence due to its potential applications in entertainment and security. Speech2Face [100] leverages a pretrained face encoder and a fixed face normalization decoder to map speech embeddings to face embeddings, but their results suffer from severe artifacts due to the decoder's limited capability. Wav2Pix [36] employs a speech-conditioned GAN architecture but fails to generalize to speakers outside its training set. Voice2Face [153] uses an identity classifier and an image discriminator to better preserve identity while generating faces from voices. However, this approach only generates fixed frontal-view faces. To disentangle factors unrelated to facial characteristics, Choi et al. [24] add an extra random latent code to implicitly cover these factors. Their work shows that vocal features pertinent to facial characteristics can be identified in latent space, as depicted in Fig. 2.4. However, their methodology only accomplishes uncondi-

tional interpolation within the identity-irrelevant space, lacking the capacity to facilitate facial animation requiring explicit control. Meishvili et al. [95] advocate that audio can help recover reasonable high-resolution faces by mixing audio with low-resolution faces. It has been demonstrated [157] that face geometry can be inferred from voices to some extent. Their studies of correlations between face geometry and voices reveal the possibility of learning face meshes from voices, and they introduce a novel Absolute Ratio Error (ARE) metric to evaluate the reconstructed 3D geometry. Nevertheless, previous studies have primarily focused on extracting appearance information from speech to generate static faces from voices.

2.1.3 Audio Emotion Recognition and Caption

Speech Emotion Recognition (SER) entails detecting and classifying emotions in spoken language, ultimately categorizing them into specific labels such as happy, sad, angry, or neutral. This field has significant applications in human-computer interaction, customer service, healthcare, and entertainment, where understanding emotional cues can enhance user experience and interaction quality. From the perspective of pattern recognition, SER [70] can be divided into three main components: feature extraction, feature selection, and feature classification. Extracted features include Mel-Frequency Cepstral Coefficients (MFCC)[32], which capture the short-term power spectrum of sound, and Linear Predictive Coding (LPC)[17], which models the vocal tract and helps in identifying the formants of speech. With advancements in deep learning, feature classifiers have evolved from methods like Linear Discriminant Analysis (LDA)[88] to neural network architectures such as CNNs and Transformers[138].

SER faces several challenges due to the inherent complexity of emotions. Emotions are often subtle and context-dependent, making them difficult to categorize accurately. For example, a person might express mixed emotions, such as feeling both happy and sad, which complicates the classification task. Additionally, cultural and individual differences can affect how emotions are expressed and perceived, adding another layer of difficulty to SER.

To tackle this challenge, employing more intricate labels such as caption descriptions instead of simple categorization labels can be beneficial. Automated audio captioning (AAC) [165, 73] stands out as a pivotal task within the au-

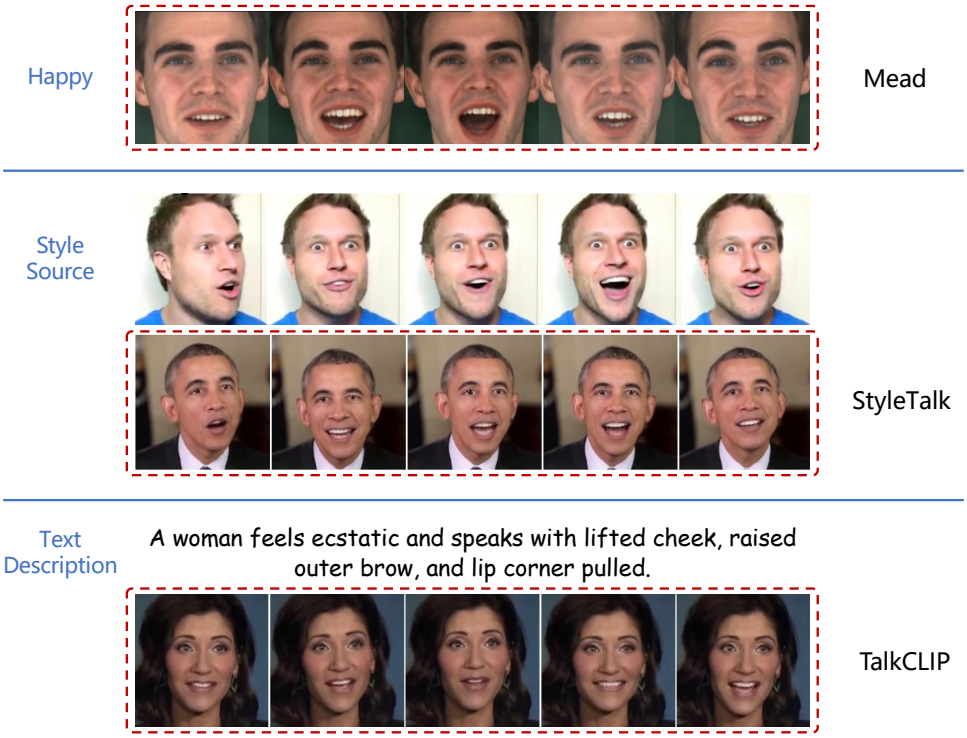

Figure 2.5: Approaches of 2D Expressive Talking Head Generation [144, 91, 90].

dio domain, articulating ambient sounds using natural language. Diverging from tasks like audio tagging [167] or sound event detection [8], AAC necessitates pinpointing specific events and articulating them fluently. Since its inception, the encoder-decoder framework [130] has emerged as the predominant solution to this challenge. Techniques such as AudioClip [50] and CLAP [160] leverage contrastive learning to forge a robust connection between audio and text, thereby enhancing the encoder-decoder paradigm. Recent research [166] has even showcased the capability of describing the emotional context of speech clips through textual means.

2.1.4 Expressive 2D Talking Face Generation

Facial expressions play a crucial role in creating lifelike talking heads [114]. Researchers [140, 158, 63, 121, 80, 65, 91, 175] endeavor to synthesize realistic facial features while ensuring accurate lip synchronization. Initially, meth-

ods [34, 45, 49, 132, 144] encoded expressions using a limited range of emotion labels represented as one-hot encodings. However, to capture nuanced facial expressions during speech, a new wave of techniques [65, 80, 91] incorporate reference videos as a richer stylistic source. As depicted in Fig.2.5, the red boxes highlight the synthesized results. The first row showcases an approach[144] relying on emotional labels for information. In contrast, the second row presents a strategy [91] utilizing another reference video as the style source.

While these methods operate effectively on RGB videos, they heavily depend on intricately designed disentanglement strategies. However, this approach frequently leads to limited expressiveness due to the inherent difficulties associated with disentanglement. Additionally, these 2D animation stylized talking face methods have restricted applicability in contexts requiring 3D representations, such as augmented reality (AR) scenarios.

Instead of requiring users to search for a stylized source, a more user-friendly approach involves directly leveraging speaking styles from the input audio. While some methods [63, 121, 164] utilize networks to extract emotion labels, their capacity is limited to inferring only a discrete number of emotion classes from audio signals. Other researchers aim to achieve rhythmic synthesis by aligning head poses [22] or expressions [169] with audio cues. However, these efforts often result in coarse alignment without adequately considering the emotional content of the audio, leading to a lack of expressiveness. To enhance the vividness and controllability of talking head generation, recent works leverage text as an interface, enabling users to specify their desired styles [90, 141], as illustrated in the last row of Fig. 2.5.

2.1.5 Speech-Driven 3D Talking Head Generation

To broaden their applicability, 3D talking head generation techniques utilize 3D representations to animate expressive facial motions. In contrast to 2D facial animation, which operates on RGB videos, 3D animation employs speech-conditioned animation, utilizing geometric representations such as 3D parametric templates [41] or implicit functions like NeRF [48] or SDF [3]. As depicted in Fig. 2.6, the left side displays some template-based samples, while the right side showcases constructed meshes from implicit functions. It's evident that the generated visualizations are more vivid with included hair details, owing to the

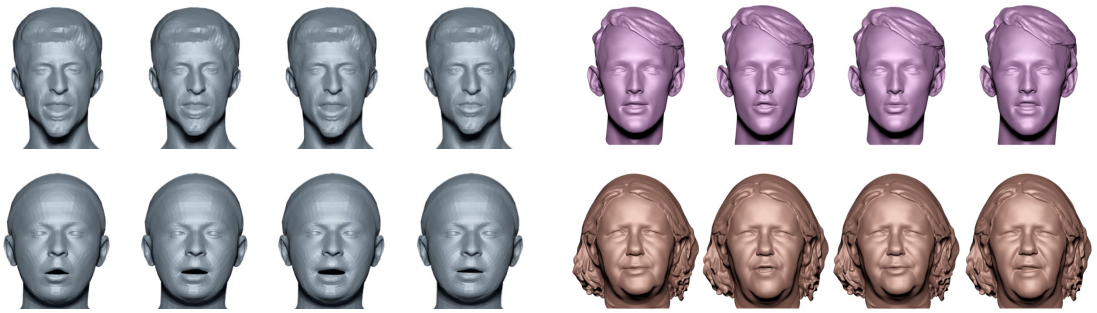

Figure 2.6: Approaches of 3D Talking Head Generation [41, 3, 113].

flexibility of implicit function representation.

In template-based speech-driven approaches, techniques like [41, 163, 31] employ speech signals to drive each vertex of a predefined mesh. While these methods effectively synchronize facial motion with the audio input, they often rely on deterministic models, leading to rigid motion in areas unrelated to speech and resulting in unnatural synthesis. To address this issue, some approaches [108] focus on capturing upper face movements by employing separate latent codes for audio-related and non-audio-related motions, such as blinking and eyebrow raises.

To overcome these limitations, recent approaches [129] have introduced a diffusion mechanism known for its remarkable generative capability. This technique yields diverse high-quality synthesis results [135, 76], incorporating more expressive facial details.

2.1.6 LLM for Cross-Modal Learning

Large language models (LLMs) have showcased profound capabilities [151, 59] as remarkable reasoning engines across various language generation tasks, credited to their emergent ability [150]. Diverse LLMs, including OPT [174], LLaMA [136], and GLM [171], can be fine-tuned or directed for various purposes [103]. Particularly, many studies aim to develop LLMs proficient in multi-modal reasoning and actioning [159], leading to the emergence of MM-LLMs. Some research indicates that LLMs might even surpass diffusion models on standard image and video generation benchmarks [5].

In the pursuit of LLMs capable of handling both multi-modal input and output,

certain approaches explore using LLMs as decision-makers [115] and employing existing off-the-shelf multi-modal encoders and decoders as tools for processing multi-modal input [187] and output [60, 166, 127].

With rich contextual prior knowledge, LLMs adeptly tackle a multitude of visual-language tasks [1, 177, 16] through appropriate prompting adaptation. Through visual instruction tuning, LLMs discern image content, reason about involved events, and generate plausible responses [78, 84, 187]. Subsequent studies [74, 125, 156] have shown that LLMs can naturally provide visual feedback by leveraging an image rendering backbone such as a diffusion model. Recently, researchers have employed LLMs to facilitate the image generation process [43, 44], where the language model exhibits impressive capabilities in layout reasoning and instruction execution.

2.2 Preliminary

This section provides an overview of the essential preliminary knowledge for this study. Firstly, we present the biological basis of vocal production, which serves the theorectical foundation of our study. Then we introduce some fundamental backbones utilized in this research, namely StyleGAN and diffusion models. Next, we discuss the foundation of parametric facial models used for representing 3D animation. Furthermore, we briefly introduce evaluation metrics for natural language generation, given their relevance to this study. Lastly, we describe the pretrained models, including SyncNet, Wav2Vec, and LLaMA, which serve as the foundational models in the proposed approaches.

2.2.1 Biological Footprint of Vocal Production

Speech is produced through a sequence of coordinated actions involving respiration, phonation, resonance, and articulation [42]. In mammals, the process begins with the lungs generating airflow that passes through the larynx. The larynx then converts this airflow into sound via the vibration of the vocal folds. This sound is subsequently filtered by the vocal tract and radiated into the environment through the lips and nostrils. This production process involves three key systems: the respiratory system, the phonatory system, and the filter system (which typically includes the resonatory and articulatory systems) as shown in

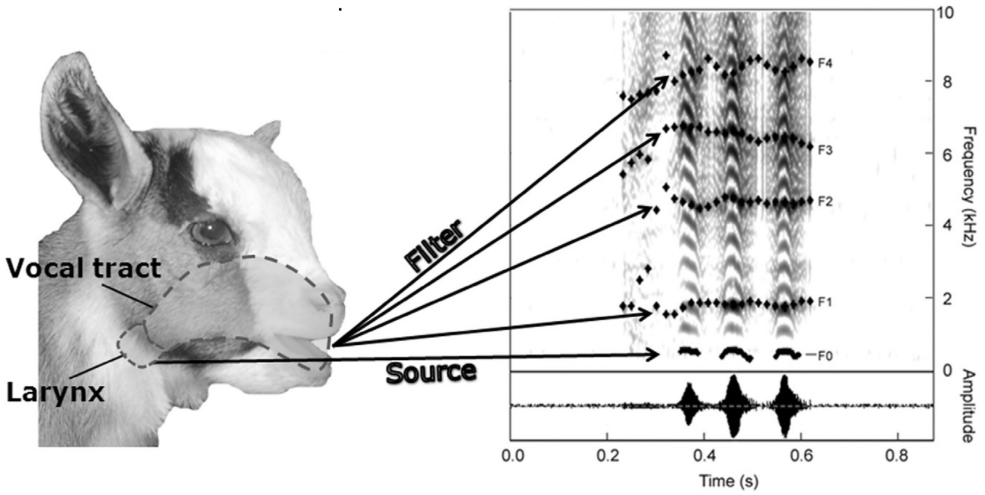

Figure 2.7: Vocal Production of Mammals [12].

	System	Location	Function in vocal production	Associated vocal parameters
Source	Respiration	Lungs and trachea	Conducting the air flow	Amplitude, duration $F0$
	Phonation	Larynx	Transforming the air into sound	$F0$
Filter	Resonance	Vocal tract	Filtering the source sound	Formats, relative engery distribution
	Arcticulation	Tongue, lips	Transforming in language speech	$F1$ and $F2$ contours

Table 2.1: Vocal production mechanism in mammals [12].

Tab. 2.1. For humans, the uniquely low position of the larynx in the throat allows us to modify the size of our oral cavity using the tongue, lips, teeth, and jaw. By constricting the vocal tract at various points, we can alter the acoustic properties of the first two formants ($F1$ and $F2$), thereby producing distinct vowel sounds. Higher formants ($F3$ and beyond) are influenced by the length of the vocal tract. This ability to fine-tune the vocal tract configuration is fundamental to the production of the wide array of speech sounds characteristic of human language.

Impact of Emotions on Vocal Parameters. Fig. 2.7 illustrates the source-filter theory of vocal production[12]. When humans experience emotions, changes occur in the somatic and autonomic nervous systems (SNS and ANS), affecting the tension and morphology of muscles involved in voice production. These modifications influence vocal parameters by altering the structure of the vocal apparatus. Specifically, the SNS primarily affects motor expression, while the ANS impacts respiration. Due to the distinct roles of the sympathetic and parasympathetic branches in different emotions, the effects of the ANS on vocalizations vary. Typically, high-arousal emotions are associated with high sympathetic tone and low parasympathetic tone. Physiological arousal mainly impacts parameters related to respiration and phonation, such as fundamental frequency ($F0$), amplitude, and duration. Emotional valence, on the other hand, is reflected in intonation patterns and voice quality, affecting the energy distribution in the spectrum. High-arousal emotions, such as fear or joy, are characterized by an increase in amplitude, $F0$, $F0$ range, and $F0$ variability. In contrast, low-arousal emotions, such as boredom, are associated with lower $F0$ and a slower speech rate.

Individual Differences on Articulation. Some studies [154] have demonstrated the differences among speakers in lingual articulation. Speech kinematic research has discovered differences between varied talkers. It revealed that tongue shapes for a sound vary widely across talkers for a single phonetic context, and continuously across the representational space. These variations highlight the complexity of speech production, as individual anatomical and physiological characteristics contribute to distinct articulatory patterns. For instance, factors such as tongue length, muscle strength, and the flexibility of different parts of the tongue can all influence how sounds are produced. Additionally, habitual speaking styles and learned speech behaviors further contribute to these articulatory differences.

2.2.2 StyleGAN

StyleGAN [68] marks a significant advancement in generative modeling, especially in creating realistic human faces. Unlike earlier GAN variants, StyleGAN introduces a novel synthesis architecture that operates across multiple levels of abstraction, enabling the generation of high-resolution images with unprecedented

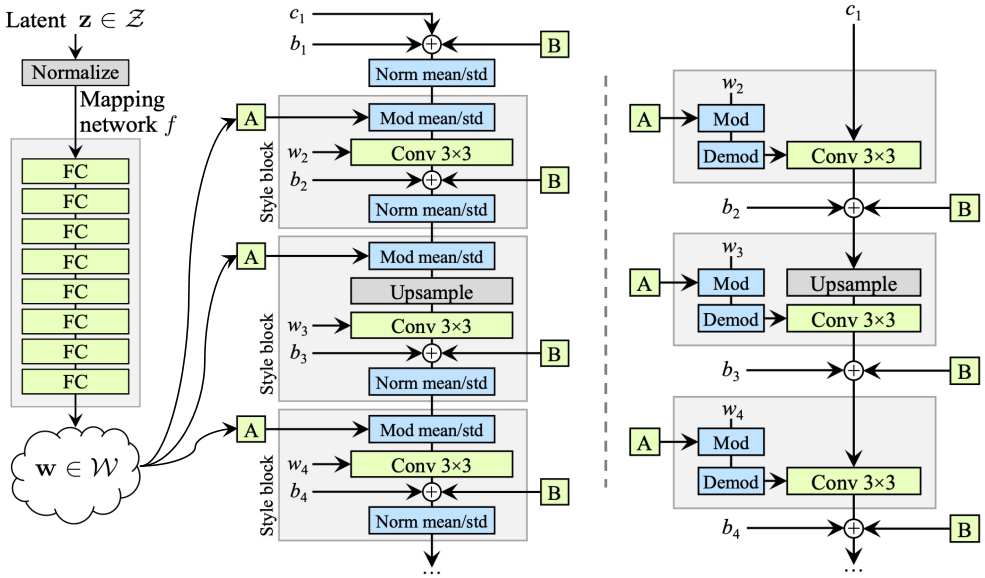

Figure 2.8: StyleGAN2 [69] (Right Side) architecture replace the instance normalization with a demodulation operation.

detail and diversity.

Central to StyleGAN is its innovative use of style-based generator and discriminator networks, which separate the latent space from the image space. This separation of $\mathbf{w}$, illustrated in Fig. 2.8, allows for finer control over various aspects of image synthesis, including pose, expression, and lighting conditions. By disentangling these factors, StyleGAN enables the creation of diverse and highly customizable images with remarkable visual fidelity.

Building on the success of its predecessor, StyleGAN2 [69] further advances the state-of-the-art in image synthesis. StyleGAN2 introduces architectural modifications such as weight demodulation, which enhance image quality and training efficiency. In each convolutional block, a multi-layer perceptron is trained to correlate random noise to a modulation vector $\mathcal{M}$ of identical dimensions to the input feature's channels. For each value w_{xyz} in the convolution kernel weight w, with x denoting its position in the input feature channels, y related to output channel numbers, and z representing spatial location, modulation and normaliza-

tion occur based on the corresponding value of x in $\mathcal{M}$:

$$w_{xyz}^{m} = \frac{\mathcal{M}_x \cdot w_{xyz}}{\sqrt{\sum_{x,z}(\mathcal{M}_x \cdot w_{xyz})^2 + \epsilon}}, \tag{2.1}$$

with the addition of a small constant ϵ to prevent numerical errors [184]. Additionally, StyleGAN2 incorporates techniques such as path length regularization, which encourages smoother and more interpretable latent space traversals, allowing for finer control over generated images.

2.2.3 Diffusion Models

The goal of generative models is to learn a distribution that approximates real data distribution $q(\mathbf{x}_0)$. The denoising diffusion probabilistic models (DDPMs) [54] present a multi-step progress to approximate $q(\mathbf{x}_0)$ with $p_\theta(\mathbf{x}_0)$ parameterized by θ, involving both a forward and reverse process.

The *forward process*, often referred to as *diffusion process*, transforms the real structured distribution into Gaussian noise, constructing a posterior distribution $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$. This process follows a Markov chain that progressively introduces Gaussian noise to the data samples. Formally,

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{2.2}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}). \tag{2.3}$$

Here, the constants β_t follow an increasing trend [54] such that when β_t approximate to 1, the x_t approximates the Gaussian noise distribution $\mathcal{N}(0, \boldsymbol{I})$.

The *reverse process*, also known as *generative process*, targets to reverse the Gaussian noise back to joint distribution $p_\theta(x_{0:T})$. Formally,

$$p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \tag{2.4}$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)). \tag{2.5}$$

Here, the variance $\Sigma_\theta(\boldsymbol{x}_t, t) = \beta_t \boldsymbol{I}$ is set as a time-dependent constant. Therefore, a generative model $\mathcal{G}_\theta$ could be devised to approximate mean value of Gaussian distribution. For conditional generation, the conditional signal $\mathbf{c}$ can be naturally

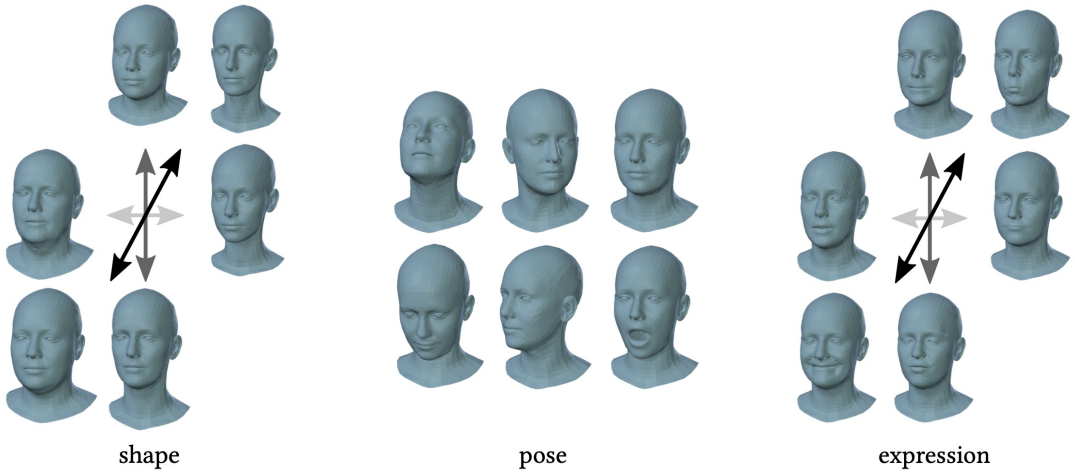

Figure 2.9: Linear Face Model of FLAME [79].

integrated into the network architecture. Formally, the model parameters θ are optimized for all sampled timestamps t and $\boldsymbol{x}$ with the following objective:

$$\mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{x},t}[\|\boldsymbol{x} - \mathcal{G}_\theta(\boldsymbol{x}, t, \boldsymbol{c})\|^2]. \tag{2.6}$$

2.2.4 Parametric Face Model

Blanz and Vetter [9] first proposes the concept of generic 3D face model, 3DMM, that learns from scanned data. 3DMM represents face model with linear bases of shape, expression and texture by leveraging Principal Component Analysis (PCA) to the collected 3D facial scans. And they release their face model as Basel Face Model (BFM) [106]. The 3DMM model is composed of shape bases $\mathbf{A}_{id}$, expression bases $\mathbf{A}_{exp}$ and mean shape $\overline{\mathbf{S}}$. A template face mesh is constructed as $\mathbf{S} = \overline{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}$ after fitting the shape and expression coefficients α. However, the BFM model is built only from head scans of 200 persons. Later approaches [11, 10] extend this face model to 10000 facial scans of more diverse subjects in neutral expression. To capture expression variation, FLAME [79] collects 3800 scans of human heads and trains a linear model within linear face shape space with articulated jaw, neck, eyballs, expressions and pose-dependent blendshapes.

As shown in Fig. 2.9, the FLAME model is a parametric 3D head model expressed as a function $\mathbf{M}(\beta, \theta, \psi) \rightarrow (\mathbf{V}, \mathbf{F})$, where the parameters consist of iden-

tity shape $\beta \in \mathbb{R}^{|\beta|}$, facial expression $\psi \in \mathbb{R}^{|\psi|}$ and pose $\theta \in \mathbb{R}^{3k+3}$ involving rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^3$. After conversion, FLAME outputs a 3D mesh with vertices $\mathbf{V} \in \mathbb{R}^{n_v \times 3}$ and faces $\mathbf{F} \in \mathbb{R}^{n_f \times 3}$, where n_v represents the number of vertices and n_f denotes the number of faces.

2.2.5 Evaluation Metrics of Text Generation

Due to the development of sequence-to-sequence deep learning technologies such as transformer-based language models, Natural Language Generation (NLG) has improved exponentially in recent years. It covers a broad range of tasks, including machine translation, summarization, question answering, dialogue generation and other open-ended generation tasks.

Evaluation Metrics. Evaluating the quality of generated text stands as a cornerstone in Natural Language Generation (NLG) research. Metrics such as BLEU [104] (Bilingual Evaluation Understudy), ROUGE [82] (Recall-Oriented Understudy for Gisting Evaluation), and METEOR (Metric for Evaluation of Translation with Explicit Ordering) serve as common tools to assess the fluency, coherence, and relevance of the generated text. Here, we provide a brief overview of the calculation process for these metrics.

The BLEU score assesses the similarity between a machine-generated translation and a reference text by measuring the precision of n-grams (contiguous sequences of n words). It computes the frequency with which the sequence of words in the output sentence appears within the reference, thereby quantifying the quality of the translation. It can be written as

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \cdot \log(p_n)\right). \tag{2.7}$$

The BLEU score calculates the weighted average of scores ranging from unigram to N-grams. The weight for each **n-gram**, denoted as w_n, typically employs a uniform weight distribution, where $w_n = \frac{1}{N}$. The p_n represents the ratio of the maximum number of times a specific n-gram occurs in the output sentence divided by the total number of candidate n-grams. Additionally, the brevity penalty (BP) is utilized to penalize shorter generations.

Formally, the p_n and BP is defined as

$$p_n = \frac{\text{Number of matching n-grams}}{\text{Total number of candidate n-grams}} \tag{2.8}$$

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-\frac{r}{c})}, & \text{if } c \leq r. \end{cases} \tag{2.9}$$

The r represents the length of the referee sentence while c is for output sequence length.

ROUGE [82] stands as another widely used metric, particularly favored for its recall-oriented perspective in tasks like text summarization and machine translation. In our work, we specifically employ its variant, ROUGE_L. This metric's calculation hinges on identifying the longest common subsequences (LCS) between the candidate and reference texts. It quantifies the overlap of words appearing in the same order in both the candidate and reference texts. For a given candidate X and reference Y, ROUGE_L is computed as follows:

$$\text{ROUGE}_L = \frac{\text{LCS}(X, Y)}{\text{Length}(Y)}. \tag{2.10}$$

The $\text{Length}(Y)$ denotes the length of the reference Y, which serves to normalize ROUGE_L scores between 0 and 1. A higher score indicates a greater overlap between the candidate and reference texts. The $LCS(X, Y)$ represents the length of the longest common subsequence between X and Y, typically calculated using dynamic programming techniques.

METEOR [7] calculates a score based on word-matching between candidate and reference sentences. It utilizes a modified F1-score formula to combine precision and recall. Formally, it is expressed as the harmonic mean of precision and recall, given by:

$$\text{METEOR} = \frac{\text{precision} \times \text{recall}}{\alpha \times \text{precision} + (1 - \alpha) \times \text{recall}}, \tag{2.11}$$

where *precision* denotes the proportion of correctly matched uni-grams in the candidate sentence, while *recall* indicates the proportion of correctly matched uni-grams in the reference sentence. The parameter α acts as a weight to balance *precision* and *recall*. This formulation takes into account both the accuracy of matching and the penalty for unmatched words in the candidate and reference sentences.

In the realm of image captioning, CIDEr [147] ((Consensus-based Image Description Evaluation)) emerges as a commonly employed metric, adept at considering both the relevance and diversity of generated descriptions. The computation of the CIDEr score typically entails the calculation of TF-IDF scores, cosine similarity, and the aggregation of similarity scores. The calculation of cosine similarity proceeds as follows:

$$\text{CosSim} = \frac{\sum_{w \in W} \text{TF-IDF}_{\text{cand}}(w) \times \text{TF-IDF}_{\text{ref}}(w)}{\sqrt{\sum_{w \in W} (\text{TF-IDF}_{\text{cand}}(w))^2} \times \sqrt{\sum_{w \in W} (\text{TF-IDF}_{\text{ref}}(w))^2}}. \tag{2.12}$$

Following the computation of TF-IDF scores, which involves operations like Term Frequency, Document Frequency, and Inverse Document Frequency Calculation, the final CIDEr score is determined by aggregating the cosine similarity scores across all n-grams. A higher CIDEr score indicates better agreement between the candidate and reference descriptions in terms of both relevance and diversity.

SPICE [2] (Semantic Propositional Image Captioning Evaluation) score stands as another metric for evaluating the quality of image captions. It assesses the quality of generated descriptions based on semantic content and structure. The SPICE score is calculated by combining both precision and recall scores, formulated as:

$$SPICE = SPICE_{content} \times SPICE_{structure}. \tag{2.13}$$

The content score, $SPICE_{content}$, evaluates the proportion of matched semantic propositions between candidate and reference descriptions, capturing the semantic similarity. Conversely, the structure score assesses the structural similarity between the candidate and reference descriptions, capturing how well the generated description follows the syntactic structure of the reference. Their detailed calculation involves parsing, matching, and scoring semantic propositions [2].

In summary, BLEU is a precision-based metric that measures the n-grams by comparing n-grams in the generated text with reference texts. METEOR tends to be more lenient and is considered better for semantic relevance. ROUGE measures recall and overlap of n-grams, focusing on how many words from the reference appear in the generated text. For the CIDEr and SPICE metrics, they are usually used for image caption evaluation. Specifically, CIDEr weights n-gram similarity by how often an n-gram occurs across multiple references, favoring less frequent but important words. The SPICE metric evaluates the quality

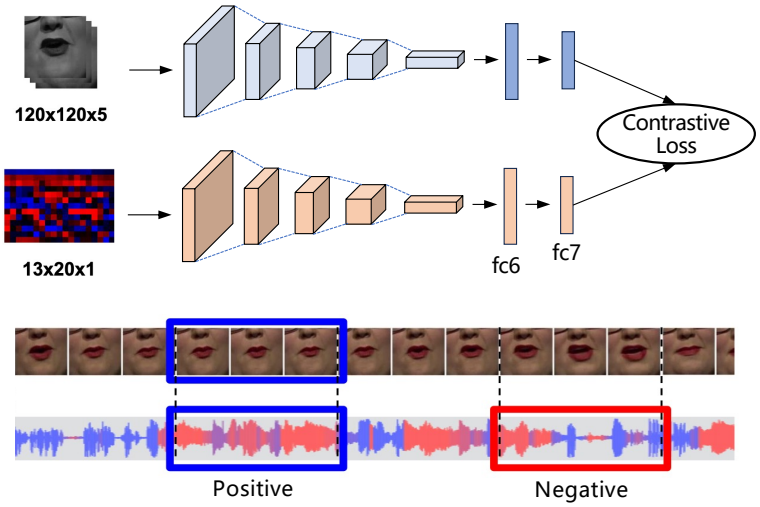

Figure 2.10: Architecture of SyncNet [28].

by converting sentences into scene graphs that capture objects, attributes and relationships, focusing on semantic correctness.

2.3 Pretrained Models

2.3.1 SyncNet

The SyncNet was first proposed by Joon Son Chung *et al.*[28] to determine the *lip-sync error* in a video. A two-stream ConvNet architecture is trained to obtain a joint embedding between the sound and mouth images from unlabeled data. The SyncNet architecture is illustrated in Fig. 2.10.

The audio is represented by a heatmap of MFCC features, where its strength indicates power at different frequency bins, and it is unrolled along the temporal dimension. For the visual modality, a sequence of grayscale images around the mouth area is utilized, typically using 5 frames of images spanning around 0.2 seconds. Both the audio and visual streams accept a spatial input of $H \times W \times C$ shape, and the network architecture adopts a stack of convolutional blocks. The training objective uses contrastive loss [25] as the optimization target, formally written as:

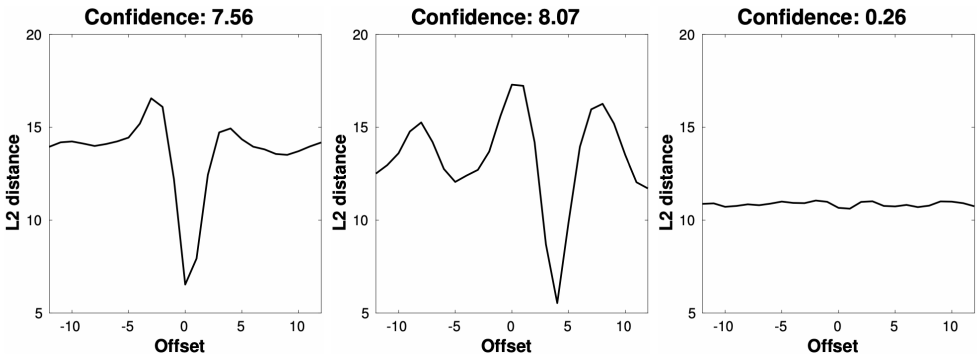

Figure 2.11: Distances measured by SyncNet [28].

$$E = \frac{1}{2N}\sum_{n=1}^{N}(y_n)d_n^2 + (1-y_n)max(T_m - d_n, 0)^2 \tag{2.14}$$

$$d_n = ||v_n - a_n||_2, \tag{2.15}$$

where the a and v are extracted fc_7 feature vectors from audio and visual streams, respectively. y_n is the binary label indicating whether audio and video inputs are from the synchronized time steps. And T_m is the loss margin. The loss calculation takes into account N positive pairs and N negative pairs. The process of obtaining *postive* and *negative* sample pairs is depicted in lower part of Fig. 2.10. Considering a video clip of 0.2 seconds, its genuine pair is chosen as the segment of speech within that duration. For its false audio-video pair, the audio from the same video is shifted by 2 seconds as a negative sample. This method ensures that the network learns temporal alignment rather than speaker differences.

With this pretrained network, audio-video synchronization can be determined by measuring the distance between their extracted features. Fig. 2.11 shows the mean distance averaged over a clip between the audio and video features given varied offsets. From left to right, three different scenarios are listed, including synchronized audio-visual situations, audio faster than video, and uncorrelated cases. Intuitively, when the L2 distance achieves its lowest value, audio and video achieve the best synchronization at that offset value. Also, the closer the distance, the better the synchronization quality. Therefore, the **Confidence** score is introduced to evaluate synchronization quality, defined as the difference between the minimum and the median of the Euclidean distances. Correspondingly, the

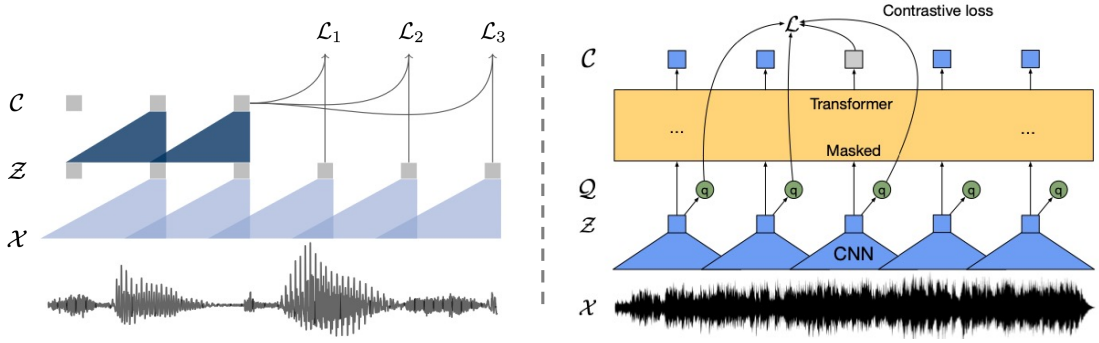

Figure 2.12: Wav2vec models [116, 6] learn a robust speech representations in an unsupervised manner.

L2 distance at this offset is defined as **LSE-D**. For scenarios where audio and visual are uncorrelated, the confidence score will be very low because there will be similar L2 distances for all time steps.

Later works [110, 184] have revealed that such pre-trained network can be used as optimization objective to improve the lip-synchronization performance, further demonstrating the effectiveness of audio-visual pre-training on unlabeled data.

2.3.2 Wav2Vec

Recent advancements in self-supervised learning of speech representations have paved the way for more efficient and effective speech processing systems. One notable approach in this domain is Wav2Vec [116, 6], which learns powerful speech representations directly from raw audio signals in an unsupervised manner. Fig. 2.12 presents the wav2vec architecture where raw audio waveform data $\mathcal{X}$ is forwarded to a stack of convolutional neural networks (CNNs): an encoder network $\mathcal{X}$ and an context network $\mathcal{C}$. The encoder network maps the audio signal to a latent space, while the context network integrates information from multiple time steps to yield contextualized representations. This model is pretrained using a causal convolutional network to predict future signals based on past samples.

During training, the model learns to distinguish a target sample $\mathbf{z}_{i+k}$, which is k steps ahead in the future, from distractor samples $\tilde{\mathbf{z}}$ drawn from a proposal distribution p_n. This discrimination is achieved by minimizing the contrastive

loss $\mathcal{L}_k$ for each step k:

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \left(\log \sigma(\mathbf{z}_{i+k}^{\top} h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} [\log \sigma(-\tilde{\mathbf{z}}^{\top} h_k(\mathbf{c}_i))] \right) \tag{2.16}$$

Here, $\sigma(x) = 1/(1 + \exp(-x))$ denotes the sigmoid function, and $\sigma(\mathbf{z}_{i+k}^{\top} h_k(\mathbf{c}_i))$ represents the probability of $\mathbf{z}_{i+k}$ being the true sample. The overall loss is computed as the sum of $\mathcal{L}_k$ across different step sizes and is applied to both the context network and encoder network to enforce learning.

In the updated version, Wav2Vec 2.0 [6], the pretext task is modified to masked modeling. The speech input within the latent space is masked, on top of which a contrastive task is defined. Meanwhile, a quantized representation $\mathcal{Q}$ is introduced to further enhance the contrastive objective as shown in right side of Fig. 2.12. In contrast, they propose to jointly learn discrete speech units and contextualized speech representations. For the context network output $\mathbf{c}_t$ of masked time step t, the model is tasked with identifying the quantized latent speech representation $\mathbf{q}_t$. This task involves navigating through a collection of $K + 1$ quantized candidate representations, denoted as $\tilde{\mathbf{q}} \in \mathbf{Q}_t$, which includes the target $\mathbf{q}_t$ and K distractors. Notably, these distractors are uniformly sampled from other time steps within the same utterance. Formally, the new contrastive loss is formuated as

$$\mathcal{L} = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)} \tag{2.17}$$

where the $sim(\mathbf{f_1}, \mathbf{f_2}) = \mathbf{f_1}^T \mathbf{f_2} / \|\mathbf{f_1}\| \|\mathbf{f_2}\|$ denotes the cosine similarity between context representations and quantized latent speech representations.

2.3.3 LLaMA

Trained on extensive corpora, Large Language Models (LLMs) have demonstrated exceptional abilities across various tasks with simple instructions [13]. A series of open-source LLMs, such as Flan-T5 [27], Vicuna [23], LLaMA [137], and Alpaca [133], have significantly advanced the field and benefited the community. Notably, LLaMA [137], which is trained solely on publicly available corpora, achieves competitive results with top-tier models like Chinchilla [56] and PaLM-540B [26], promoting the democratization and study of LLMs.

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2.2: Architecture details of LLaMA [137].

LLaMA models are available in sizes ranging from 7 billion to 65 billion parameters, accommodating various computational budgets. The LLaMA network utilizes the standard transformer architecture [138], with detailed architecture and training hyperparameters summarized in Table 2.2. To enhance performance, several techniques have been implemented, building on designs from previous models like GPT-3, PaLM, and GPT-Neo. Specifically, the LLaMA network employs input pre-normalization via RMSNorm [173], following GPT-3's approach. It replaces the original ReLU activation function with the more advanced SwiGLU [118], as seen in PaLM. Additionally, rotary positional embedding (RoPE) [124] is used at each transformer layer, replacing the standard absolute positional embeddings used in GPT-Neo.

Network Architectures. The overall architecture of LLaMA [137] adopts a transformer network [138], as illustrated on the left side of Fig.2.13. Unlike RNN-based approaches such as LSTM[55], transformers handle sequential data in parallel, enabling them to model long-term dependencies more effectively. The transformer follows an encoder-decoder structure connected by an attention mechanism.

The encoder processes an input sequence $\mathbf{X} = (x_1, \cdots, x_n)$ and learns a sequence of latent embeddings $\mathbf{z} = (z_1, \cdots, z_n)$ using N stacks of self-attention blocks and fully connected layers. The decoder then generates the output sequence $\mathbf{Y} = (y_1, \cdots, y_m)$ sequentially, based on these latent representations, using N stacks of blocks. Unlike the encoder, each decoder block includes an additional layer that performs feature fusion of the encoder's output and a masked sequence input via multi-head attention. Here the n and m denote the lengths of the input and output sequences, respectively.

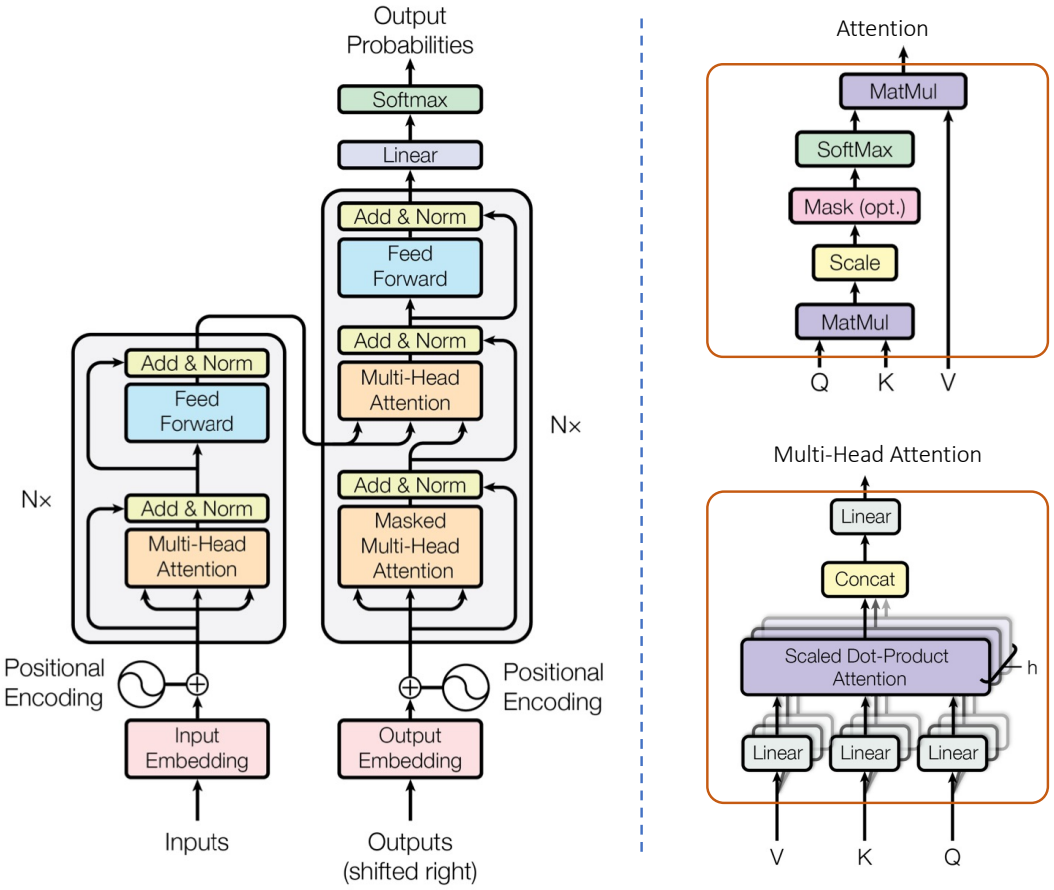

Figure 2.13: Basic transformer architecture [138].

A crucial component of this architecture is the multi-head attention mechanism, depicted on the right side of Fig. 2.13. Instead of performing a single attention operation, multi-head attention utilizes multiple parallel attention layers that project the queries, keys, and values to dimensions d_k, d_k and d_v, respectively. Such practice allows the model to jointly attend to information from different representation subspaces for a set of Q, K and V vectors with d_{model} dimensions, thereby enhancing learning and performance. Formally,

$$\text{MultiHead}(Q, K, V) = Concat(head_1, \cdots, head_h)W^O \tag{2.18}$$

$$\text{where } head_i = Attention(QW_i^Q, KW_i^k, VW_i^V). \tag{2.19}$$

The projection matrices W are defined as $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}, W_i^K \in \mathbb{R}^{d_{model} \times d_k}, W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. The h indicates the number of scaled dot-

product attention operation involved within a MultiHead module. Its calculation process is illustrated on top-right side in Fig. 2.13. Formally, the Q, K, Vs are integrated with

$$\text{Attention}(Q, K, V) = \text{softmax}\frac{QK^T}{\sqrt{d_i}}V. \tag{2.20}$$

3 Research Proposal

3.1 Realistic Talking Face Synthesis

Given a clip of human speech, this task targets to synthesize a talking face that well synchronized with the speaker voice. Below, we summarize several crucial aspects that a realistic talking face synthesis system is expected to encompass:

- **High Image Quality**: Low-resolution images or renderings with evident blurring or generative artifacts will significantly degrade manifestation performance.
- **Lip Synchronization**: Mouth movements must be well synchronized with the speech content.
- **Vivid Demonstration**: The system should encompass distinct talking aspects such as natural head poses, alterations in emotion, and speaker characteristics.
- **Consistent Synthesis**: Synthesis results should accurately reflect the speaker's facial characteristics or speaking status.
- **Versatile Control**: The synthesis system will be more flexible and user-friendly with a greater range of controllable perspectives.

Meeting these requirements proves particularly challenging due to the complex cross-modality nature involving various interleaved elements. The synthesis system must adeptly integrate disparate knowledge from both audio and visual sources. The difficulty of realistic generation arises from two perspectives:

1) Initially, the input information is intertwined, with certain weak vocal features presenting uncertainty and ambiguity. Consider the human voice, which

encompasses speech content, emotional status, and speaker identity, all interwoven. While speech content can be easily discerned, emotional status and speaker identity may lack obvious clues, presenting various plausible possibilities. Effectively extracting the required information from human speech within a unified framework poses non-trivial challenges.

2) Conversely, integrating these pieces of information for talking face synthesis demands cross-modality consistency across various aspects, such as facial appearance and motions. This task is further complicated by the multitude of movement patterns within a talking face, including head pose variations, facial action units, and mouth movements. Head movements, akin to ridge movements, possess fewer degrees of freedom (DoF), whereas lip motions exhibit frequent changes corresponding to spoken words, contrasting with the low-frequency alterations in emotional expressions and other facial actions.

Numerous studies have delved into enhancing the realism of talking face synthesis.

1) In pursuit of improving visual quality, several approaches have been proposed to augment synthesis regions [175, 184], ensure temporal consistency [168], or enhance synthesis resolution [176]. However, many of these methods overlook the importance of maintaining identity consistency between human speech and speaker identity, potentially impacting the audience's experience.

2) To enhance motion naturalness, previous works either model the correlation between dynamic head poses and audio rhythm [22, 129] or incorporate external representations [91, 90, 108] such as emotion labels or video clips as style references during generation. The former approaches offer limited expressive head dynamics, resulting in coarse alignment that neglects emotional nuances present in voice. Meanwhile, the latter studies necessitate manual style source selection by users, leading to unnatural applications. Few works exploit direct utilization of the underlying style information conveyed by human speech in talking face synthesis.

Despite the ambiguous nature of identity features and speaking status within vocal features, accurately capturing this fine-grained information and modeling it significantly contributes to natural and realistic talking face synthesis. We argue that a core pathway to achieving such synthesis lies in *devising a sophisticated strategy to fully exploit the fine-grained cross-modal associations*. Developing a system that integrates the aforementioned ambiguous cues and accurately reflects

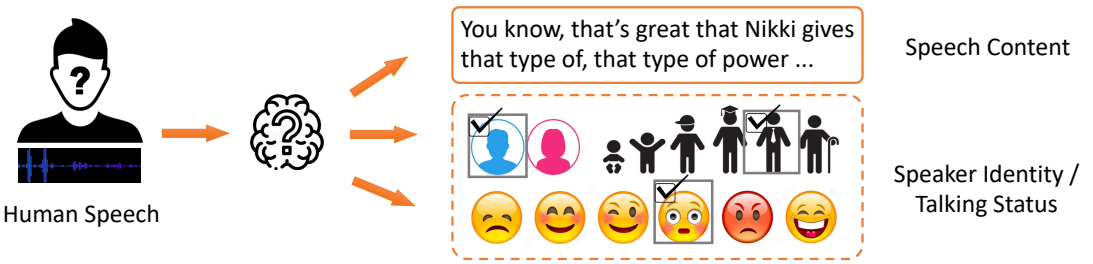

Figure 3.1: Talking face system that also considers speaker identity and status.

them onto faces demands meticulous architectural design and training paradigms, representing an open problem worthy of exploration.

3.2 Research Approach

We target to accomplish realistic talking face synthesis via enhancing fine-grained cross-modal association. Our key insight is to *fully leverage the underlying information present in human voice and incorporate it into the generative process*. Recognizing that previous talking face synthesis systems often overlook *speaker identity and speaking status* in human speech, we focus on uncovering and leveraging these factors during the synthesis of a talking face as shown in Fig. 3.1. In the following sections, we present our proposed strategy of audio-visual association learning and its derived novel applications to validate the effectiveness of this approach.

3.2.1 Audio-Visual Association Learning

As mentioned above, the difficulty of realistic synthesis lies in both recognition and exploitation of relevant information in generation process. We introduce contrastive learning and disentanglement strategy to resolve these two issues, respectively.

Feature Enhancement via Contrastive Learning. Leveraging relevant features from human speech involves first identifying and enhancing features related to speaker identity and speaking status. These features, however, are often weak or ambiguous compared to speech content. Contrastive learning [52] offers an ef-

fective approach to enhancing feature extraction by maximizing mutual information across modalities. Previous studies [20, 110, 75, 181, 188] have demonstrated the benefits of enhancing mutual information between audio and visual modalities for lip synchronization. Inspired by this, we construct relevant pairs with similar information to enhance the extraction of speaker identity and talking status within audio. For speaker identity extraction, we leverage video frameworks that capture strong personal characteristics such as facial appearance and shape. As talking status expression is more implicit, we propose leveraging explicit language descriptions to enforce it. In summary, we borrow explicit visual cues to enforce the feature representation to human voice. To maximize the mutual information, we need to carefully construct paired features $\mathbf{F}_{ex}^{v} \in \mathbb{R}^{l_c}$ and $\mathbf{F}_{im}^{a} \in \mathbb{R}^{l_c}$ as positive samples and choose N^{-} negative audio features $\mathbf{F}_{im}^{a-} \in \mathbb{R}^{N^{-} \times l_c}$. Following the InfoNCE [101], the loss function can be formally represented as

$$\mathcal{L}_c = -\log[\frac{\exp(\mathcal{D}(\mathbf{F}_{ex}^{v}, \mathbf{F}_{im}^{a}))}{\exp(\mathcal{D}(\mathbf{F}_{ex}^{v}, \mathbf{F}_{im}^{a})) + \sum_{j=1}^{N^{-}} \exp(\mathcal{D}(\mathbf{F}_{ex}^{v}, \mathbf{F}_{im(j)}^{a-}))}]. \tag{3.1}$$

The $\mathcal{D}$ indicates feature distances measurement, where the cosine distance $\mathcal{D}(\mathbf{F}_1, \mathbf{F}_2) = \frac{\mathbf{F}_1^{\mathbf{T}} * \mathbf{F}_2}{|\mathbf{F}_1| \cdot |\mathbf{F}_2|}$ is commonly adopted and closer features often render larger scores.

Feature Integration within Disentangled Space. Realistic synthesis necessitates the natural integration of disparate aspects of generation, including personal identity, head movement, lip sync, and facial action units. With the enhanced features, the talking face synthesis system is expected to integrate them with other information from audio-visual sources. Therefore, an approach to appropriately incorporate them into the corresponding aspects precisely is crucial for successful manifestation. We propose accomplishing latent space disentanglement of the generative network before integration. Subsequently, the utilization of enhanced features becomes a problem of learning an effective mapping into the corresponding semantic space. Such incorporation practice ensures that the generation progress is well delegated to each specific semantic space in advance, preventing undesirable interference with other elements. Specifically, we aim to prevent speaker identity or talking status from impacting the lip movement of our synthesized face. Therefore, the generative model is expected to *complementarily disentangle into speech content space and identity space or talking status space*, respectively.

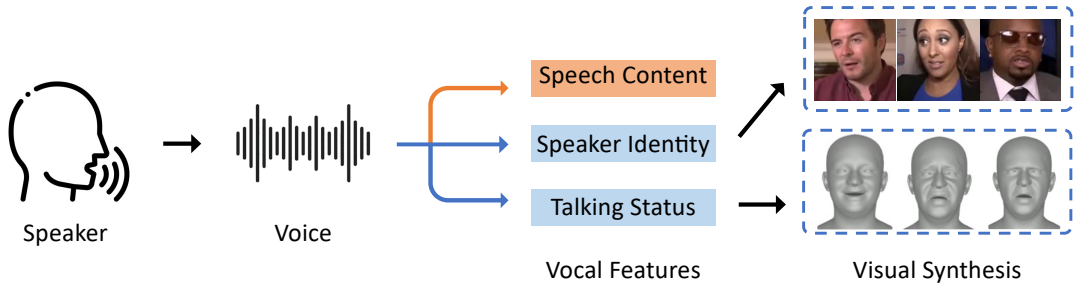

Figure 3.2: Applications of Realistic Talking Face via Cross-Model Associations.

3.2.2 Case Studies on Novel Talking Applications

By strengthening the cross-modal connections between different components of the system, we can explicitly synthesize fine-grained characteristics. This enhancement not only improves the realism of talking face synthesis but also enables innovative applications as shown in Fig. 3.2. Here, we outline specific association strategies, such as leveraging speaker identity and speech status from human speech, and their resulting novel applications:

- **Speech2Talking-Face** [126]. This novel approach involves directly inferring a talking face from human speech by disentangling associations between facial images and the underlying identity characteristics within the voice.
- **AVI-Talking** [128]. This system functions as a talking face instruction tool, allowing users to manipulate the speaker's talking status. Multi-Modal Large Language Models meticulously model the underlying talking status within human speech and facial movements.

As depicted in Fig. 3.2, human voice carries abundant information, with speech content typically being explicit while speaker identity and talking status features are often weak and ambiguous. In our research, we focus on enhancing these subtly vocal attributes and incorporating them into our generative framework. The Speech2TalkingFace work specifically amplifies speaker identity, resulting in a talking system that is attuned to individual characteristics. Meanwhile, the AVI-Talking work reinforces implicit talking status, fostering the development of audio-visual instruction synthesis systems.

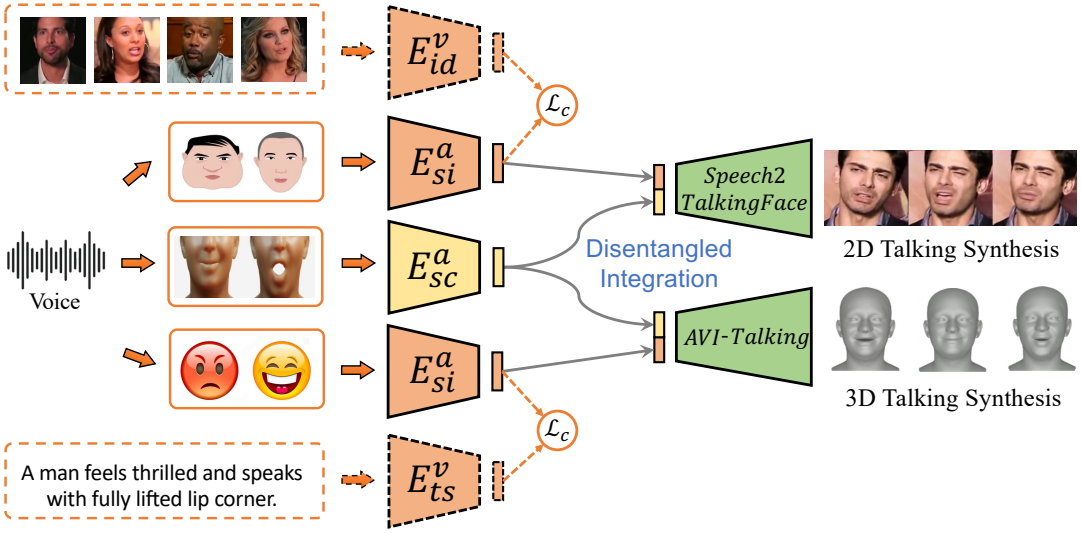

Figure 3.3: System Overview.

3.3 System Overview

Fig. 3.3 illustrates the overview of our proposed system. In order to fully leverage the rich information embedded in human voice, we have incorporated several strategies, such as contrastive learning, to amplify the vocal features. These implicit features, encompassing speaker identity and speech status, are meticulously integrated into the synthesis network, enabling innovative applications as depicted on the right side of Fig. 3.3. Speaker identity derived from the voice is reinforced and, in conjunction with speech content, is separately input into the disentangled space of a generator, thereby advancing the Speech2TalkingFace application. Similarly, the speech status is extracted and linked to the disentangled latent space for AVI-Talking. Importantly, our association learning strategies exhibit versatility, as they are applicable to both 2D and 3D talking face synthesis, as demonstrated, confirming their broad utility.

4 Inferring and Driving a Face with Synchronized Audio-Visual Representation

In this section, we present *Speech2Talking-Face (S2TF)*, a unified framework that achieves face inferring and driving simultaneously from audio. The pipeline of our method is illustrated in Fig. 4.1. Below we first introduce the overview of our framework, then we explain the synchronization between audio and visual representations in two latent spaces. Finally, we show the procedures of our visually assisted curriculum learning for visual generation.

4.1 Task Formulation

Normally the problem of face reconstruction from voice is formulated in a way of conditional image generation. Given a clip of speech A (transformed to spectrogram S), the ideal result would be an image I' with the same identity (appearance) as the source person. In our setting, rather than generating an image, we propose to generate a clip of video $V' = \{I'_1 \ldots, I'_K\}$ which should be close to the video $V = \{I_1, \ldots, I_K\}$ which produces the speech A. The generated frames should not only have the visual appearance of the source identity but also have accurate lip motion synced with audio.

4.2 Proposed Approach

4.2.1 Framework Overview

However, it is difficult to build supervision for both appearance and lip motion at the same time. To this end, we adopt a similar framework as [14] with a

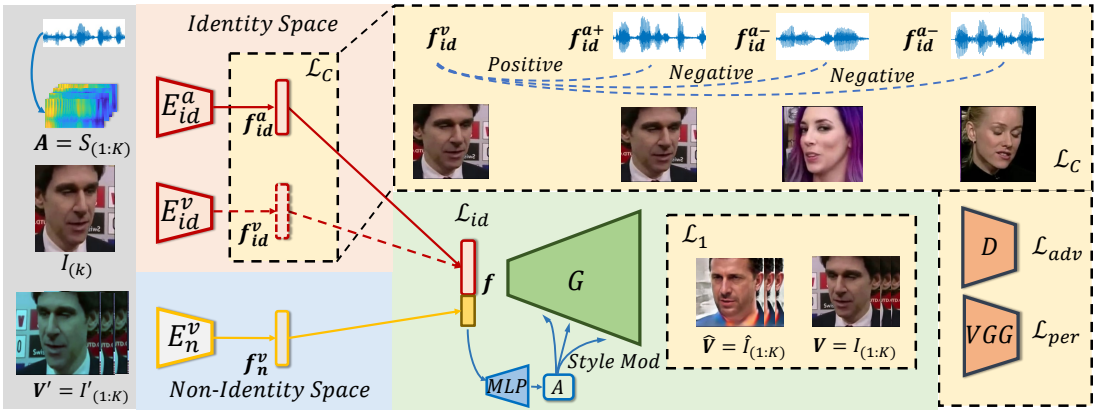

Figure 4.1: The overview of our proposed Speech2Talking-Face pipeline.

style-based generator [69]. The idea is to explicitly encode frame I_k to an identity space and a frame I_i^+ with data augmentation to an identity-irrelevant space. The style-based generator is then leveraged to reconstruct the original frame I_i. The data augmentation aims at preserving the pose and expression information while changing the texture and shape which are highly correlated with identity. The style-based generator enables separate control of both the information required through the training of frame reconstruction.

Based on this framework, our goal is to encode the speech information to both the *identity* and *identity-irrelevant* latent spaces, where they can be mapped to the image domain by the generator.

4.2.2 Disentangled Pretraining via Explicit Visual Cues

Due to the ambiguity nature of implicit speaker identity information in human voice, it is quite challenging to directly train an end-to-end network. We target to first pretrain a network where the speaker identity and explicit speech content are separately represented so that these two types of information do not inference with each other. One natural way to achieve this is devising a framework composed of one decoder and two encoders. One encoder targets to capture facial appearance while another one accounts for facial movements. A decoder is required to integrate information from both of them. Following the typical face reenactment procedure, this problem is modeled as a cross-frame reconstruction task. Formally, Given a K-frame video clip $\boldsymbol{V} = \{I_{(1)}, \ldots, I_{(K)}\}$, the natural

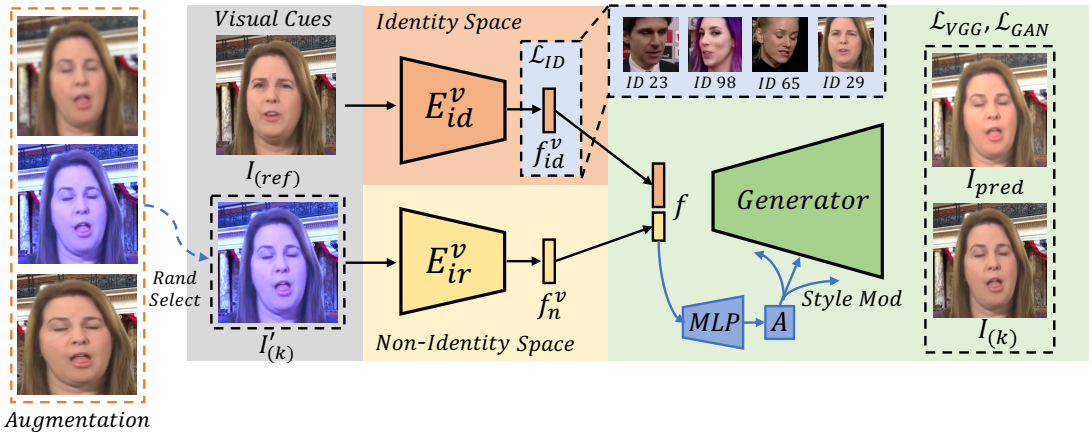

Figure 4.2: Formulation of pre-trained disentangled space via explicit visual cues.

training goal is to generate any *target* frame $I_{(k)}$ conditioned on one frame of identity reference $I_{(ref)}$ ($ref \in [1, \ldots, K]$). For *target* frame $I_{(k)}$ synthesis, it is necessary to include facial motion and head pose information at target time step k. But directly incorporate I_k frame as input will cause the copy-paste issue where the network could simply learn to output the input.

It has been verified that data augmentation becomes an effective strategy to identify distinct kinds of information from images [14, 80, 184]. The basic principle is to utilize data augmentation to imitate data variation. For instance, the perspective transformation and color transfer can be applied to an image to cover the facial appearance changes. Therefore, augmented target frames $\boldsymbol{V}' = \{I'_{(1)}, \ldots, I'_{(K)}\}$ are leveraged to provide the facial motion and head pose information. To this end, the disentangled framework with a cross-frame reconstruction objective is determined. Overall, this design choice follows previous face reenactment approach [14]. In the pursuit of better disentanglement, we also utilize perspective transformation following [184]. As depicted in Fig. 4.2, the training objective is to reconstruct the *target* frame $I_{(k)}$ with VGG Loss $\mathcal{L}_{VGG}$ and Discriminator Loss $\mathcal{L}_{GAN}$.

In contrast to previous approach that utilizes AdaIN [61] for feature injection, we introduce style modulation strategy [69] in our generator. At every convolutional block, a multi-layer perceptron is trained to correlate random noise to a modulation vector $\mathcal{M}$ of identical dimensions to the input feature's channels. For each value w_{xyz} in the convolution kernel weight w, with x denoting its po-

sition in the input feature channels, y related to output channel numbers, and z representing spatial location, modulation and normalization occur based on the corresponding value of x in $\mathcal{M}$:

$$w_{xyz}^{m} = \frac{\mathcal{M}_x \cdot w_{xyz}}{\sqrt{\sum_{x,z}(\mathcal{M}_x \cdot w_{xyz})^2 + \epsilon}}, \tag{4.1}$$

with the addition of a small constant ϵ to prevent numerical errors [184].

4.2.3 Audio-Visual Representation Synchronization

The solution is to synchronize the audio-visual representation within two spaces from the same source with contrastive learning [28, 181].

Identity Space Learning. The learning of the *identity* space is of more importance given that the question of recovering appearances from voices is still not well-solved. We name the visual-to-identity encoder E_{id}^{v}, which encodes feature f_{id}^{v} from the image I_k. Then we leverage a speech encoder E_{id}^{a}, for encoding the audio identity feature f_{id}^{a} from speech S.

We first build positive and negative pairs from different talking face videos and their corresponding speeches. As the speech embedding f_{id}^{a} should be close to the image embedding f_{id}^{v}, image embeddings from other videos $f_{id}^{v-}(i)$ can be served as negative samples that our speech embedding should repel. We use the term $\cos\theta_{(f_1,f_2)} = \frac{f_1^{\mathbf{T}} * f_2}{|f_1| \cdot |f_2|}$ to define the cosine distance between two features f_1 and f_2. Then the contrastive loss for the identity information synchronization can be written as:

$$\mathcal{L}_{sync}^{id} = -\log[\frac{e^{\cos\theta_{(f_{id}^{v}, f_{id}^{a})}}}{e^{\cos\theta_{(f_{id}^{v}, f_{id}^{a})}} + \sum_{i=1}^{M} e^{\cos\theta_{(f_{id}^{v-}(i), f_{id}^{a})}}}], \tag{4.2}$$

Where M is the number of negative samples in one batch.

Additionally, as our model is trained on the VoxCeleb2 [30] dataset, identity labels are used for more compact identity representation learning. Normally a fully connected (fc) layer with weight W would be leveraged. Then a softmax cross-entropy loss is served for classification.

It has been verified that after the convergence of a classification task, the weight W_j for each class j actually lies in the centroid of all feature embeddings

from identity j. As a result, we share the W for both audio and visual so that given a same identity, the features from both modalities will be encoded around a same centroid. Further more, we would expect the encoded features to be more distinctive, thus the angular-based loss function ArcFace [35] is leveraged for any f_{id} from both audio and visual modality. The idea is to bring closer the features within one class and their centroid and repel other centroids. The loss function can be written as:

$$\mathcal{L}_c = -\sum_{i=1}^{M} \log[\frac{e^{s\cos(\theta_{(W_j,f_{id})}+m)}}{e^{s\cos(\theta_{(W_j,f_{id})}+m)} + \sum_{p\neq j}^{P} e^{s\cos\theta_{(W_p,f_{id})}}}], \tag{4.3}$$

where s is a scale factor and m is a margin. The loss form is quite similar to the contrastive loss in Eq. 4.2.

Speech Content Subspace Learning. The *identity-irrelevant* space is encoded by encoder E^v_{ir} in the visual domain through data augmentation, where an image I_i^+ would be encoded to feature $f_{ir}^{v_i}$. Thus the audio spectrogram S is divided to $S = \{s_1, \ldots, s_K\}$ to match the frame numbers with a sliding window. Normally, we would expect the speech features $f_{ir}^{a_i}$ encoded from speech encoder E^a_s to be the same as the visual features $f_{ir}^{v_i}$. However, the *identity-irrelevant* space contains information such as head pose, which cannot be inferred from speech. Instead of synchronizing f_s^a with f_{ir}^v, we focus more on learning the dynamic changes between frames, which includes mainly speech content information.

Specifically, we select the input frame to the identity spaces I_k as the first frame I_1, and send it also to E^v_{ir} to obtain the feature $f_{ir}^{v_1}$. Then we compute the changes of the rest of the frames as $\Delta\mathbf{f}_{ir}^v = \{f_{ir}^{v_2} - f_{ir}^{v_1}, \ldots, f_{ir}^{v_K} - f_{ir}^{v_1}\}$. The encoded speech features $\mathbf{f}_s^a = \{f_s^{a_2}, \ldots, f_s^{a_K}\}$ are thus synchronized with $\Delta\mathbf{f}_{ir}^v$ through contrastive loss:

$$\mathcal{L}^s_{sync} = -\log[\frac{e^{\cos\theta_{(\Delta\mathbf{f}_{ir}^v,\mathbf{f}_s^a)}}}{e^{\cos\theta_{(\Delta\mathbf{f}_{ir}^v,\mathbf{f}_s^a)}} + \sum_{i=1}^{M} e^{\cos\theta_{(\Delta\mathbf{f}_{ir}^{v-(i)},\mathbf{f}_s^a)}}}]. \tag{4.4}$$

The formulation of the negative pair is the same as [28].

After the learning procedure, we obtain a speech feature f_s^a which lies also in the *identity-irrelevant* space. However, as only content information can be preserved in the speech feature, we identify that these speech features actually formulate a *speech content* subspace within the *identity-irrelevant* space.

4.2.4 Training Paradigm

We adopt a curriculum training paradigm to enable convergence and information balancing within this framework. The idea is to leverage network training in the visual domain as a probe to guide the learning of audio information.

Visual Domain Training. Having obtained the identity feature f_{id}^{v} and identity-irrelevant feature f_{ir}^{v} from the visual embedding, the style-based generator [69] G targets to decode the concatenated feature $f^{v} = \text{cat}(f_{id}^{v}, f_{ir}^{v})$ to a face image $I^{v\prime}$. Detailedly, the feature is sent through MLPs and served to modulate the weights of the generator's convolution kernels. Then the weights are normalized through the demodulation operation. Such design has been proven to be effective in general image generation tasks. The reconstructed image can be written as $I^{v\prime} = \text{G}(f^{v})$. At this stage, the identity encoder is supervised by the angular loss written in Eq. 4.3.

The training of the generator relies on the image reconstruction loss $\mathcal{L}_{rec}$ and adversarial loss $\mathcal{L}_{adv}$. For image reconstruction, we employ $\mathcal{L}_1$ loss in RGB space, and the perceptual loss which constrains high-level features. For any generated image I':

$$\mathcal{L}_{\text{per}} = \frac{1}{L}\sum_{l=1}^{L} \|\phi_l(I) - \phi_l(I')\|_1 \tag{4.5}$$

$$\mathcal{L}_{rec} = \mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per}. \tag{4.6}$$

Where ϕ_l denotes the lth layer of a pretrained VGG19 [119] network. For adversarial loss, we apply a multi-scale discriminator [148] D with N_D layers. The adversarial loss can be written as

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \min_{G}\max_{D} \sum_{n=1}^{N_D} (&\mathbb{E}_I[\log D_n(I)] \\ &+ \mathbb{E}_{\mathbf{I'}}[\log(1 - D_n(I')]).\end{aligned} \tag{4.7}$$

Therefore, the overall training objective is written as

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{per} + \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_{adv}, \tag{4.8}$$

where the λ_1 and λ_2 are balancing coefficients.

Audio to Visual Synchronization Training. At the second stage, we attempt to synchronize audio features to visual features. Thus the visual identity encoder E^{v}_{id} and visual identity-irrelevant encoder E^{v}_{ir} are fixed.

The audio identity encoder E^{a}_{id} is trained on the shared W with E^{v}_{id}. It is also supervised by the angular loss (Eq. 4.3), and the contrastive loss (Eq. 4.2). While the audio speech content encoder E^{a}_{s} is trained under the supervision of Eq. 4.4. During this stage, we also adopt the curriculum training strategy that first selects videos of different identities as negative samples, then moving to the videos of the sample identity but speaking different contents.

Final Finetuning. At last we gradually finetune all models. Particularly, a feature of $f^{va} = \mathrm{cat}(f^{v}_{id}, f^{a}_{s}+f^{v_1}_{ir})$ is firstly formed and finetuned with the generator G to reconstruct $I^{va\prime} = \mathrm{G}(f^{va})$, where $f^{v_1}_{ir}$ provides the initial pose information. This is for the generator to adapt to the speech content information while keeping the identity unchanged. Specifically, if both identity and content are sourced from speech, the reconstruction loss (Eq. 4.6) would emphasize only on the appearance of the face but neglect the mouth shapes.

Finally the audio features are combined together as $f^{a} = \mathrm{cat}(f^{a}_{id}, f^{a}_{s} + f^{v_1}_{ir})$ to generate $I^{a\prime} = \mathrm{G}(f^{a})$, and fintuned with all models. This procedure is also crucial given that the strong expressive ability of style-based generator would implicitly balance the information between the two features. The final cost function can be written as:

$$\mathcal{L}_{all} = \mathcal{L}_{\mathrm{adv}} + \lambda_r \mathcal{L}_{\mathrm{rec}} + \lambda_c \mathcal{L}_c + \lambda^{id}_{sync} \mathcal{L}^{id}_{sync} + \lambda^{s}_{sync} \mathcal{L}^{s}_{sync} \tag{4.9}$$

where the λs weight different loss terms.

Please be noted that during testing, the feature f^{a} for reconstruction can be derived in two ways. The first is to encode an arbitrary $f^{\hat{v}}_{ir}$ from any face image $\hat{I}$ to form $f^{a} = \mathrm{cat}(f^{a}_{id}, f^{\hat{v}}_{ir} + f^{a}_{s})$. In this way, our generated results will preserve the initial pose of the image $\hat{I}$. On the other hand, we could simply omit this pose guidance feature and direct leverage $f^{a} = \mathrm{cat}(f^{a}_{id}, f^{a}_{s})$. Frontalized faces can thus be generated under such scenario.

4.3 Experiments

4.3.1 Experimental Settings

Dataset. We use the popular in-the-wild dataset VoxCeleb2 [30] in our experiments. It contains a total of 6,112 celebrities. 5994 speakers are selected for training and 118 speakers are selected for testing. Images are extremely diverse in this dataset, containing different light conditions, large head poses, and varied video quality.

Implementation Details. We adopt the encoder from [181] as E^{v}_{ir}. The structure of visual identity encoder E^{v}_{id} is ResNeXt50 [162] and the audio encoders are ResNetSE34. Our generator G consists of 6 blocks of style modulated convolutions. The length of identity features f^{a}_{id} and f^{v}_{id} are set as 2048 while identity-irrelevant features f^{v}_{ir} and f^{a}_{s} are set as 512.

We conduct our experiments using PyTorch deep learning framework with eight 16 GB Tesla V100 GPUs. Images are cropped to 224×224. The audio inputs are mel-spectrograms processed with FFT window size 1280, hop length 160 with 80 Mel filter-banks. A clip of human voice lasting 3.2 seconds is used for our speech-to-identity mapping. During testing, We retrieve an arbitrary image from other identities as the pose source for our generated identity. All λs in loss functions are empirically set to 1.

Evaluation Protocol. Since prior research has not addressed the challenge of simultaneously generating facial appearance and lip motion from speech within a single model, we introduce two evaluation metrics to comprehensively assess our approach. These metrics evaluate the quality of voice-to-face mapping and the accuracy of lip movements. In Section 4.3.2, we gauge the consistency of speaker identity between the synthesized image and the human voice by comparing with existing approaches such as [153, 100] which focus on voice-to-face synthesis. Meanwhile, in Section 4.3.3, we assess dynamic lip motion accuracy by comparing with methodologies like [21, 110], which concentrate on lip-synchronization.

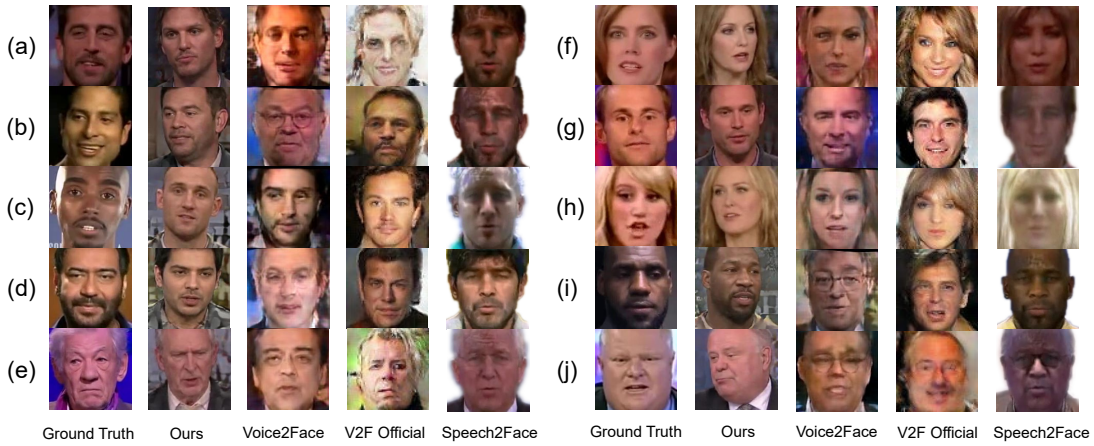

Figure 4.3: Qualitative comparison of our model and previous methods.

4.3.2 Voice-to-Face Mapping Evaluation

In this section, we focus on the evaluations of voice-to-face mapping where the relationship between human voice and static generation is considered.

We compare our model with state-of-the-art method voice2face [153]. The generated image resolution in their official release is 64×64. For a fair comparison, we also re-implement their method to produce 128×128 facial images on the same dataset as ours. We also re-implement Speech2Face [100] as described in their paper. The test set of VoxCeleb2 is leveraged in our evaluation, where all identities are not seen during training.

Qualitative Results on Voice-to-Face Generation. As illustrated in Fig. 4.3, images generated by Voice2Face and V2F official contain unstable backgrounds and unnatural artifacts. While results from Speech2Face are blurry with indistinct facial appearances, the quality of our generated faces is comparable to real reference facial images. Moreover, we can observe that our faces synthesized from different voices are age- and gender-matched with the speaker. This proves that our model can capture more underlying relationships between vocal and physical properties. Furthermore, with the help of the pose reference feature, our generated faces have more diverse poses, which expands our application in real-world scenarios.

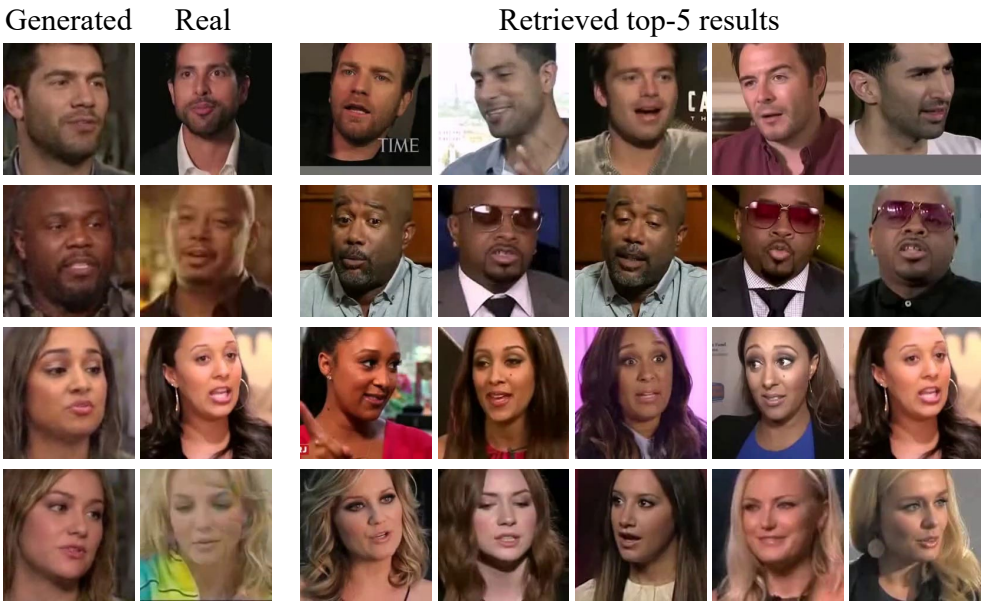

Figure 4.4: Top-5 retrieved images.

Quantitative Results on Voice-to-Face Generation. Similar to previous work [100], metrics on similarity, retrieval performance and image quality are adopted for evaluation. The similarity is measured by the cosine and L1 distances between the synthesized face and its corresponding ground truth embedded by a pre-trained FaceNet [117]. Retrieval performance is validated by querying the dataset with a produced image, targeting to retrieve the ground truth speaker image. We calculate R@K, indicating the ratio of successful retrievals in the top-K similar faces. Image quality is validated by VGGFace score (VFS), which is proposed to replace the Inception Score in face settings. We also conduct the ablation study on whether our centroid learning procedure is effective, thus our model without it (S2TF w/o CL) is also evaluated on these metrics.

The results are listed in Table 4.1. Our model achieves higher similarity scores, and higher retrieval ratios than other methods, suggesting that our model generates more similar facial images with ground truth. Some examples of the top-5 retrieval results are visualized in Fig. 4.4. It can be seen that our model could retrieve images with similar facial attributes as real images, verifying that our identity encoder successfully captures the audio-visual association. Additionally, our model achieves the highest VFS among all methods, which we attribute to

	Similarity		Retrieval			Quality
Method	cosine ↑	L1 ↓	R@1 ↑	R@5 ↑	R@10 ↑	VFS ↑
Rand	-	-	1.00	2.00	5.00	-
Speech2Face	0.304	65.34	8.43	16.21	24.05	32.38±1.92
Voice2Face	0.269	67.60	7.49	14.02	19.87	34.74±2.32
S2TF w/o CL	0.382	63.54	9.49	19.87	28.75	38.30±3.12
S2TF	**0.397**	**60.84**	**10.65**	**22.85**	**30.80**	**39.00±2.90**

Table 4.1: The quantitative results on VoxCeleb2 in embedding similarity, retrieval @K, and VGGFace Score.

Method(AUC)	Rand ↑	G ↑	N ↑	A ↑	GNA ↑
Rand	50.0	50.0	50.0	50.0	50.0
PIN	**78.5**	61.1	**77.2**	**74.9**	58.8
S2TF w/o CL	74.8	68.6	71.6	72.3	63.6
S2TF	76.0	**69.2**	75.2	73.1	**64.0**

Table 4.2: Cross-modal matching under varying demographics.

the expressiveness of style-modulation.

Voice-Face Association Matching. We conduct an experiment on voice-face matching with PIN [97] which is specifically designed for this task. We use standard metrics, area under the ROC curve (AUC), in speaker verification following their work. The model is evaluated by sampling negative test pairs while holding constant each of the following demographic criteria: gender (G), nationality (N), and age (A).

The results are listed in Table 4.2. Though our method is not designed for cross-modal matching, our results are comparable to theirs. Particularly, we outperform their results on gender and the average of the three aspects.

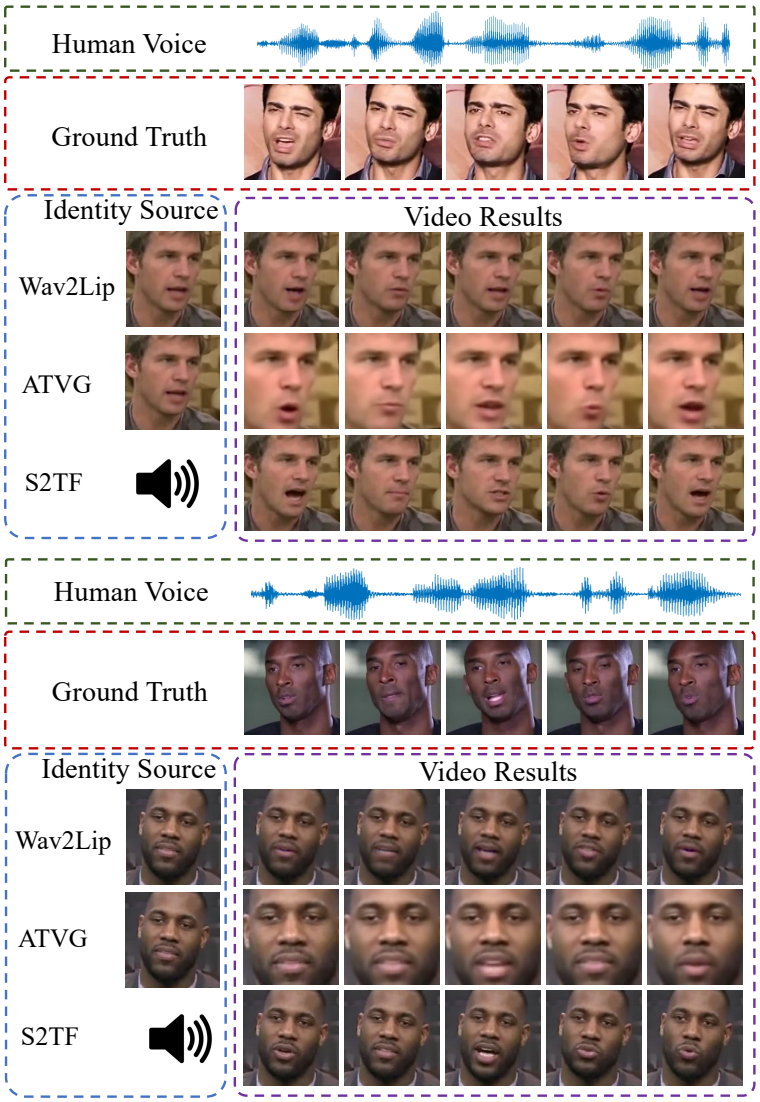

Figure 4.5: Comparison with talking-face baselines.

4.3.3 Lip Motion Evaluation

As we can directly generate dynamic talking faces from audio, we also show the lip sync results. Please be noted that our goal can also be achieved by combining voice-to-face reconstruction and audio-driven talking face generation. However, such a pipeline is too complex and non-flexible. Here we send our generated one frame to the talking face model of ATVG [21] and Wav2Lip [110].

Method	LMD ↓	Sync_{conf} ↑
ATVG	**6.49**	4.5
Wav2Lip	12.26	4.3
S2TF	6.88	**5.7**

Table 4.3: Quantitative results of lip synchronization.

Evaluation Results. It can be seen from Fig. 4.5 that while ATVG has lower qualities, the lip motion of Wav2Lip is not as accurate as ours. In Table 4.3 shows the LMD landmark distance, and the confidence score from SyncNet [28]. It can be seen that we also achieve comparable results as the methods specifically designed for this area. Particularly, if we change f_{id}^{a} to any f_{id}^{v} of an existing frame, we can directly generate talking face videos with this framework.

4.3.4 Additional Evaluation

Effect of Varied Audio Length. Different audio length will contain different amounts of identity-related and irrelevant information, thus it necessary to decide an appropriate audio length. As is presented in Table 4.4, 3.2-second audio length achieves the best performance for both similarity and retrieval metrics. This may be caused by the fact that sufficient information is already included in 3.2 seconds audio. Thus, increasing audio length may merely lead to model difficulties and worse performance. Example images produced with different audio length are shown in Fig. 4.6. We can find that increasing audio length to 4.8 leads to inaccurate facial characteristic.

Visualization of Audio-Visual Latent Representation. To study the latent representation of speaker identity, we further plot their t-SNE visualization as depicted in Fig. 4.7. We can observe that the male and female identities are clustered closely for both audio and visual representation, which explains why our approach obtains superior results in gender class as shown in Tab. 4.2. Overall, we could find that the visual embedding is separated better than the voice embedding. Such result is also conformed to our intuition because facial appearance is easier to be captured by visual images compared to human voice.

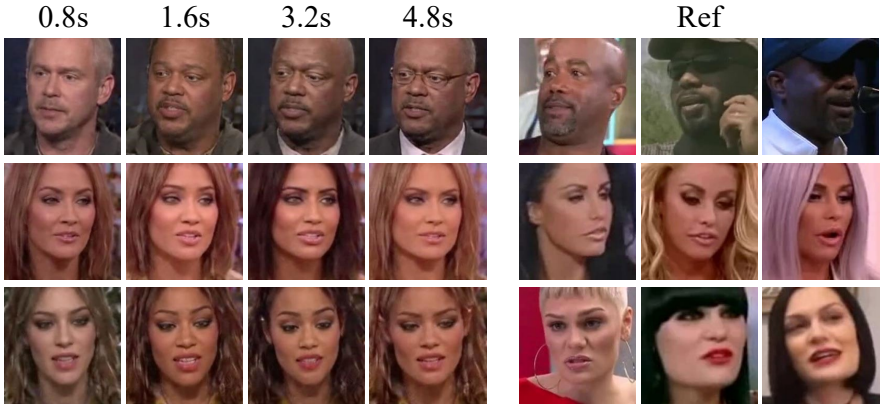

Figure 4.6: Generated facial images under different audio length setting and their corresponding reference images.

Audio	Similarity		Retrieval			Quality
Length(s)	cosine ↑	L1 ↓	R@1 ↑	R@5 ↑	R@10 ↑	VFS ↑
0.8	0.370	62.24	8.41	18.12	25.06	36.29±3.66
1.6	0.392	61.02	9.74	20.36	28.74	38.68±2.59
3.2	**0.397**	**60.84**	**10.65**	**22.85**	**30.80**	**39.00±2.90**
4.8	0.344	63.76	6.88	16.51	23.07	25.83±3.27

Table 4.4: The quantitative results on VoxCeleb2 with different utilized audio length.

Ablation Studies. Apart from the evaluation of S2TF without centroid learning above, we have also experimented on other types of contrastive losses in Eq. 4.2 such as the traditional L_2 form or cosine form. Their influences on the final results are subtle. However, using the softmax form (Info NCE) stables the training process. The model cannot render reasonable results if there is no contrastive loss.

As for the curriculum paradigm, the visual domain pretraining is essential for the network to balance the identity and identity-irrelevant information. Intuitively, reconstructing a face from visual features is definitely easier than from audio features.

User Study. We conduct a user study to further evaluate all methods. We

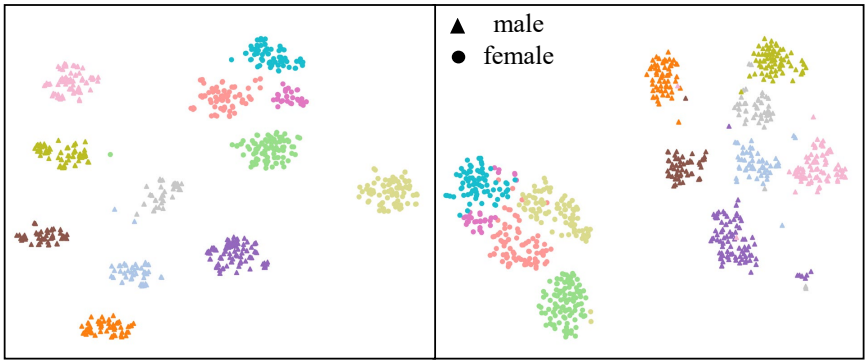

Figure 4.7: t-SNE visualizations of audio and visual embedding in latent space.

Method	Similarity ↑	Quality ↑
Speech2Face	3.76	3.64
Voice2Face	3.12	3.27
S2TF	**3.95**	**4.32**

Table 4.5: User study measured by Mean Opinion Scores.

randomly select 40 generated images from different identities in the test split of VoxCeleb2 [30]. A total of 18 participants are asked to rate the image quality and similarity between the synthesized image and its reference with the widely used mean Opinion Scores (MOS) rating protocol (from 1 to 5). The results are summarized in Table 4.5. We can see that our model achieves the highest MOS on both metrics. Particularly, our approach outperforms Speech2Face [100] and Voice2Face [153] on image quality by a large margin.

Derived Applications. Benefited by the disentangled design of our style-based generator, our framework is also capable of supporting more sophisticated applications. As demonstrated in Fig. 4.8, our framework could accept another video as pose source to guide the generation process. Intuitively, the human voice has both speaker identity and speech content information. Thus a talking video with frontal face could be synthesized. As for synthesizing faces with arbitrary head poses, we could simply leverage addition operation in content-irrelevant space. Specifically, the visual encoder extract the feature from a reference image and

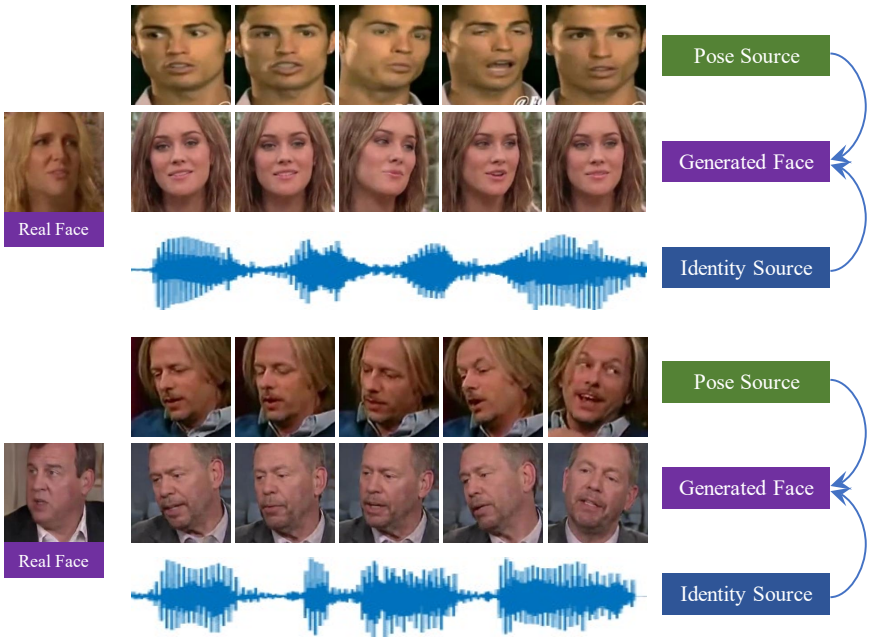

Figure 4.8: Pose control application benefited by disentangled design.

add it to the speech content feature, then the pose information is injected to our system.

4.4 Related Work

Driving a face to talk with either a clip of audio [21, 189, 181, 110, 122, 186, 64, 22, 183] or video [155, 170, 14, 72] has long been the topic of interest for both the fields of computer graphics and artificial intelligence. Earlier methods focus more on modeling a specific person [131, 72]. Recent studies propose to achieve arbitrary subject or one-shot talking face generation driven by audios or videos. A number of studies borrow intermediate representations such as landmarks [170, 21] into this task. But this is not convenient in our settings, as voice-reconstructed faces have no fixed landmarks. Certain landmark-free methods [181, 110, 183] rely on skip-connections thus require an image as network input.

Specifically, Burkov *et.al.* [14] disentangles facial identities and poses in a style-based framework. However, they require a meta-learning stage on ground truth videos. Our work leverages this framework's advantages, offering a streamlined latent space where encoder-decoder connections become unnecessary. This facili-

tates a seamless transition from explicit visual cues to speech input. Furthermore, the well-documented statistical association between human voice and facial appearance [98, 97, 71, 57] provides a theoretical basis for cross-modal learning. Thus, we find the concept of audio-visual association learning feasible and employ it as the basis for our curriculum training strategy.

For convenience, we have listed the related works of Speech2Talking-Face (S2TF) in Table 4.6. Notably, our work stands out as the only approach that integrates speech identity-based generation, lip movements, and pose control within a single unified framework.

	Year	Speech Identity	Lip Movements	Pose Control
Voice2Face [153]	2019	✓	✗	✗
Speech2Face [100]	2019	✓	✗	✗
PIN [97]	2018	✓	✗	✗
ATVG [21]	2019	✗	✓	✗
Wav2Lip [110]	2020	✗	✓	✗
PD-FGC [142]	2023	✗	✓	✓
S2TF	2021	✓	✓	✓

Table 4.6: Relation of Speech2Talking-Face with State-of-the-Art Methods

4.5 Summary

In this paper, we propose *Speech2Talking-Face (S2TF)* framework, which aims at inferring and driving a face from speech in a unified framework. We emphasize the properties of our method: **1)** Through our explicit design of two synchronized spaces, we successfully generate high-quality facial appearances with accurate lip sync from speech only. **2)** Our synchronization in the identity space learns a more compact identity association between speech and human faces. **3)** As a unified framework, our model has the capacity to go beyond its designed purpose. For example, the task of audio-driven talking face generation can be directly handled.

5 Learning Audio-Visual Instructions for Expressive 3D Talking Face Generation

5.1 Task Formulation

In this section, we aim to fully *leverage the speaker's talking status information from their voice*. Our objective is to animate a template mesh with synchronized lip movements and consistent facial expressions from an audio clip. Specifically, our system takes an input speech $\mathbf{A}^{1:T_a} = \{\mathbf{a}_i\}_{i=1}^{T_a}$ and generates a time series of FLAME parametric coefficients $\xi^{1:T_a} = \{\theta_i, \psi_i\}_{i=1}^{T_a}$. These include the facial expression coefficients $\psi \in \mathbb{R}^{|\psi|}$ and pose coefficients $\theta \in \mathbb{R}^{3k+3}$. We do not predict the identity coefficients $\beta \in \mathbb{R}^{|\beta|}$, as our primary focus is on facial expressions. For more details about the FLAME parametric face model, please refer to Sec. 2.2.4.

5.2 Proposed Approach

5.2.1 Framework Overview

Instead of directly learning to synthesize a talking face from speech, we propose integrating facial movements descriptions to facilitate realistic generation. As illustrated in Figure. 5.1, our framework, **AVI-Talking**, comprises two main stages: an audio-visual instruction stage and a talking face synthesis stage connected by visual instructions of detailed facial expression descriptions. Given a clip of speaker speech $\{a_i\}_{i=1}^{T_a}$, it is first processed by Large Language Models (LLMs) to propose visual instructions encompassing plausible facial detail descriptions. Subsequently, these visual instructions, together with audio clip, are

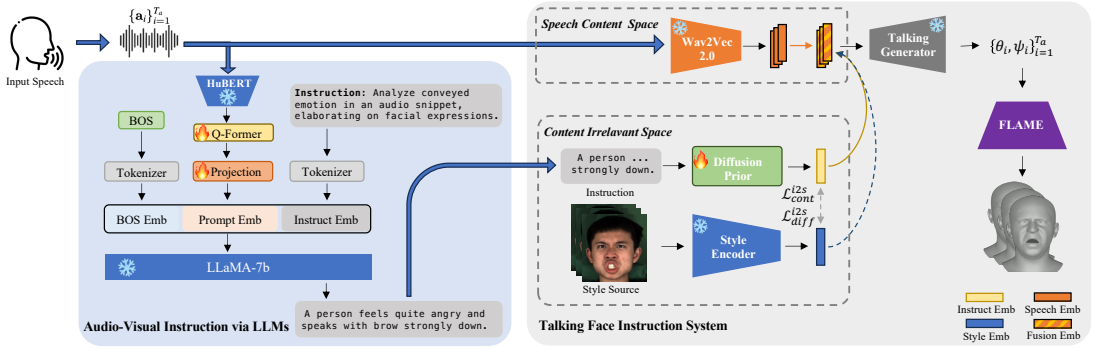

Figure 5.1: The overall pipeline of our Audio-Visual Instruction Talking (AVI-Talking) Framework.

separately fed into the talking face instruction system to generate a time series of 3D parametric coefficients $\{\theta_i, \psi_i\}_{i=1}^{T_a}$.

5.2.2 Disentangled Expressive Motion Prior

As depicted in Figure. 5.2, we target to establish a disentangled latent space, where the speech content related lip-movements and facial expressions correlated with speaking state are distinctly represented in *speech content* space and *content irrelevant* space, respectively. Concretely, in speech content space we employ a pretrained ASR network, Wav2Vec 2.0 [6] to encode the speaker audio $\mathbf{A}^{1:T_a}$. These extracted speech features capture semantic content information, which is subsequently utilized by the talking generator for syllable pronunciation. In order to encode additional talking style information, we point out the existence of *content irrelevant* space for representing content-repelling information such as talking styles, poses and speaker identity.

To learn the *content irrelevant* space, we employ a transformer-based style encoder [91] designed to capture content-repelling information. For a given talking video, we randomly select S reference frames to serve as the source for the speaking state. These frames are then processed by the FLAME model to obtain coefficients $\{\theta_s, \psi_s\}_{s=1}^{S}$, where the coefficient at time t is excluded. Subsequently, these coefficients are fed into the style encoder to extract a comprehensive speaking state representation for the video. To successfully predict coefficients at the current time step $\{\theta_t, \psi_t\}$, we rely on both the speech feature $\mathbf{A}^t$ in the *speech*

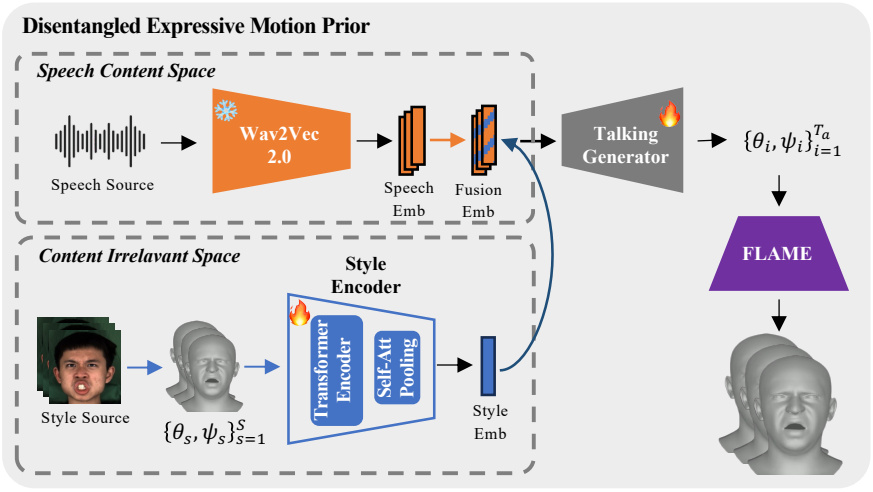

Figure 5.2: Disentangled expressive motion prior is separated to two complementary latent spaces, *speech content* space and *content irrelevant* space.

content space and the extracted style information in the *content irrelevant* space. The complementary nature of these properties naturally facilitates the learning of disentangled spaces.

5.2.3 Audio-Visual Instruction via LLMs

As illustrated on the left side of Figure 5.1, the audio-visual instruction module takes a time series of a speaker's audio clip as input and aims to generate an instruction sentence describing detailed facial movements that conveys the individual's speaking state. The key is to *develop a prompting strategy to effectively leverage the rich contextual prior knowledge inherent in LLMs.*

Specifically, we leverage a pre-trained LLaMA as our base text generation model. In order to comprehend the speaker's speaking status existing in audio modality, the audio signal needs to be projected into text embedding of language model. Due to the success of pretrained-model such as HuBERT [58] on Speech Emotion Recognition [96] (SER) tasks, we leverage HuBERT to encode the audio signal. Subsequently, a typical Q-Former [78, 1] architecture is employed to aggregate and extract speaking style information, bridging the gap between acoustic feature and visual facial descriptions. Concretely, the Q-Former architecture leverages the standard Perceiver network [1] to compress speech input to

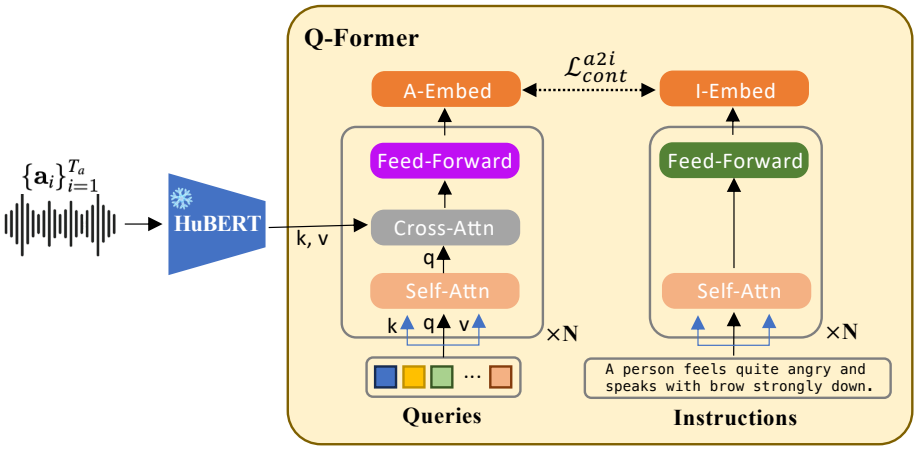

Figure 5.3: The Q-Former architecture.

a fixed-length audio embedding $\mathbf{F}_{si}^{a} \in \mathcal{R}^{q_a \times l}$. A contrastive loss $\mathcal{L}_{cont}^{a2i}$ is applied to encourage the queries extract audio representation that are most relevant to visual instructions. A linear projection layer is then learned to map the aligned feature to language model's input space. Combining the "BOS" (Beginning-of-Sequence) token with the instruction embedding, the audio prompt embedding is fed to LLaMA to prompt plausible expressive facial movements consistent with speaker status. Note that the instruction embedding is obtained by tokenizing the pre-defined instruction templates. In our experiment, we utilize instruction sentences like *Analyze conveyed emotion in an audio snippet, elaborating on facial expressions*. We manually craft 10 sentences with similar meanings and randomly sample one during the training phase.

Speech Feature Compression via Learnable Queries. The audio features extracted from HuBERT encapsulate complex information, including speech content, emotional status, and acoustic details. To effectively prompt the language model, it's essential to first comprehend and extract relevant facial movement information from the speech. Here, we employ the Q-Former architecture [78, 1] to achieve this task.

As depicted in left side of Figure 5.3, learnable queries with fixed length are utilized to aggregate and compress speech information by cross-attention. Notably, such practice results in an audio embedding $\mathbf{F}_{si}^{a} \in \mathcal{R}^{q_a \times l}$ with the same dimensionality as the query length q_a. This design choice simplifies the learning process

and enhances generalization performance when handling speech inputs of varying lengths. Subsequently, the audio embedding is fed to a projection module for prompt embedding in language model space. To implement this, we fine-tune a small number of parameters in the input projection layers for domain adaptation.

Contrastive Audio-Visual Instruction Alignment. To eliminate unnecessary information such as speech content, environment noise and focus on extracting facial movements related feature, we adopt contrastive learning [101] protocol to constrain the output of learned queries $\mathbf{F}^a_{si} \in \mathcal{R}^{q_a \times l}$. The contrastive learning paradigm aligns audio embeddings and instruction features to maximize their mutual information. This is achieved by enhancing higher audio-instruction similarity of positive pairs against those of negative pairs. Specifically, we feed the corresponding instruction through a text transformer and obtain an instruction embedding as shown in the right side of Figure 5.3. Its output embedding of $[CLS]$ token is $\mathbf{F}^i_{si} \in \mathcal{R}^l$. Since there are q_a query embeddings, we average $\mathbf{F}^a_{si}$ across all queries to obtain the $\bar{\mathbf{F}}^a_{si} \in \mathcal{R}^l$ and apply contrastive learning as follows:

$$\mathcal{L}^{a2i}_{cont} = -\log[\frac{\exp(\mathcal{D}(\bar{\mathbf{F}}^a_{si}, \mathbf{F}^i_{si}))}{\exp(\mathcal{D}(\bar{\mathbf{F}}^a_{si}, \mathbf{F}^i_{si})) + \sum_{j=1}^{N^-} \exp(\mathcal{D}(\bar{\mathbf{F}}^a_{si}, \mathbf{F}^{i-}_{si(j)}))}]. \tag{5.1}$$

The paired in-batch samples are regarded as positive samples $(\bar{\mathbf{F}}^a_{si}, \mathbf{F}^i_{si})$ while the unpaired N^- samples are taken as negative samples $(\bar{\mathbf{F}}^a_{si}, \mathbf{F}^{i-}_{si(j)})$. Here we opt for cosine distance $\mathcal{D}(\mathbf{F}_1, \mathbf{F}_2) = \frac{\mathbf{F}_1^{\mathbf{T}} * \mathbf{F}_2}{|\mathbf{F}_1| \cdot |\mathbf{F}_2|}$ as feature distance measurement.

Instruction Generation via Projection Layer Finetuning. After the Q-Former is pre-trained to contrastively align acoustic features to visual facial descriptions. Subsequently, the Q-Former is frozen, and we fine-tune the input linear projection layer of LLaMA-7b to achieve visual instruction prediction as shown in Figure 5.1. Specifically, We follow the general text generation training paradigm [187] to learn this projection layer.

5.2.4 Instruction-Following Talking Face Synthesis

With the obtained facial instructions, a talking face synthesis network aims to animate a mesh template with synchronized lip movements and expressions as

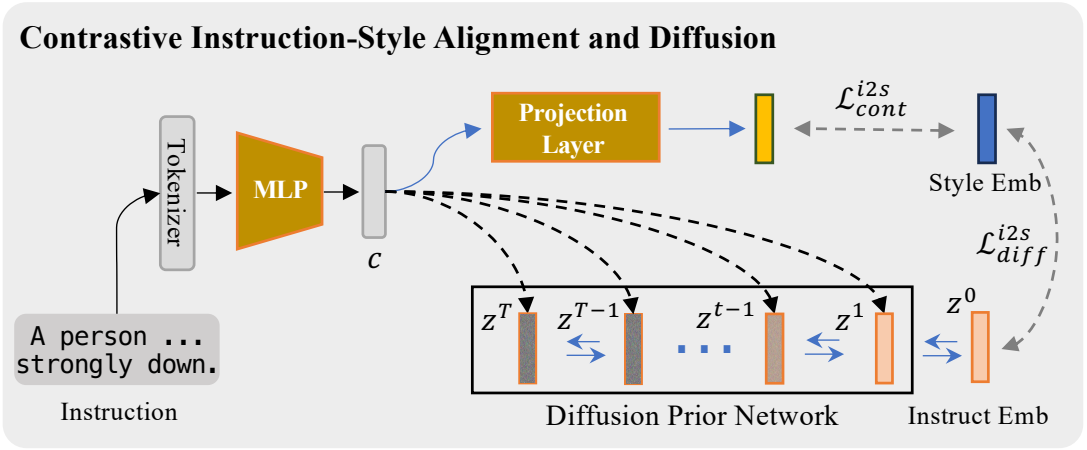

Figure 5.4: Diffusion within the content irrelevant space.

illustrated on the right side of Figure 5.1. The movements of the lips and facial expressions exhibit a high degree of correlation with each other [91]. For example, specific pronunciations often convey relevant emotions. To address this correlation and potential entanglement, we propose initially training a disentangled talking prior [113, 34], wherein the speech content space and content irrelevant space are distinguished (shown in Figure 5.2). Subsequently, a diffusion prior module (shown in Figure 5.4) is devised to bridge the gap between instruction text and talking styles within the identified content irrelevant space.

Contrastive Instruction-Style Alignment and Diffusion. Once the content irrelevant space is identified, a natural way for cross-modality generation is to map visual instruction to the representation within this space [90]. As depicted in Figure 5.4, we contrastively align the visual instruction with style embedding to obtain a aligned feature c, upon which a diffusion prior network is employed to further map it towards the distribution of the pre-trained talking prior. At first, a Multi-Layer-Perceptron (MLP) network is derived to first align latent instruction representation with style embedding. The typical contrastive loss $\mathcal{L}_{cont}^{i2s}$ is employed, following standard CLIP training process [111]. Formally,

$$\mathcal{L}_{cont}^{i2s} = -\log[\frac{\exp(\mathcal{D}(\mathbf{F}_{ci}^{i}, \mathbf{F}_{ci}^{s}))}{\exp(\mathcal{D}(\mathbf{F}_{ci}^{i}, \mathbf{F}_{ci}^{s})) + \sum_{j=1}^{N^-} \exp(\mathcal{D}(\mathbf{F}_{ci}^{i}, \mathbf{F}_{ci(j)}^{s-}))}]. \tag{5.2}$$

The $\mathbf{F}_{ci}^{i}$ indicates the content-irrelevant instruction feature, which is obtained by

passing the aligned latent instruction representation c through a projection layer. The $\mathbf{F}_{ci}^{s}$ denotes its corresponding embedded style feature within the content-irrelevant space. The batch-wise $(\mathbf{F}_{ci}^{i}, \mathbf{F}_{ci}^{s})$ instruction and style feature pairs are taken as positive samples while the unpaired N^{-} instances $(\mathbf{F}_{ci}^{i}, \mathbf{F}_{ci(j)}^{s-})$ are considered as negatives samples. Similarly, we adopt cosine distance $\mathcal{D}(\mathbf{F}_1, \mathbf{F}_2) = \frac{\mathbf{F}_1^{\mathbf{T}} * \mathbf{F}_2}{|\mathbf{F}_1| \cdot |\mathbf{F}_2|}$ as feature distance measurement.

However, since this multi-modal contrastive learning strategy only pushes the instruction embeddings to hold close direction with their associated style image features, which is prone to cause disjoint embeddings due to the existance of modality gap [81]. To further activate motion prior that expects visual style embeddings, we introduce a diffusion prior network to bridge the modality gap by mapping to their distributions.

For the diffusion prior network $\mathcal{F}_\theta$, we leverage the typical decoder-only Transformer architecture to iteratively predict the denoised style embedding $\mathbf{z}^t$ conditioned on the above representation $\mathbf{c}$. Instead of imposing error prediction formulation [54], we directly train the network to predict unnoised style embedding $\mathbf{z}$ from noised embedding $\mathbf{z}^t$ sampled at time step t. Formally,

$$\mathcal{L}_{diff}^{i2s} = \mathbb{E}_{\mathbf{z},t}[\|\mathbf{z} - \mathcal{F}_\theta(\mathbf{z}, t, \mathbf{c})\|^2] \tag{5.3}$$

where we apply the naive Mean-Square Error (**MSE**) to the prediction result.

Therefore, the overall learning objective of visual instructions to speaking styles generation can be written as

$$\mathcal{L}^{i2s} = \mathcal{L}_{cont}^{i2s} + \lambda^{i2s}\mathcal{L}_{diff}^{i2s}, \tag{5.4}$$

where λ is the balancing coefficient. In our experiment, we empirically set it to 30, following a similar protocol to previous work [111].

5.3 Experiments

5.3.1 Experimental Settings

Datasets. We train both audio-visual instruction module and talking face instruction network on MeadText [90] dataset. Evaluation is conducted on test set of RAVEDESS [89] and MeadText. Since both datasets are made of RGB videos,

we obtain reconstruction results by Emoca [33] and render the facial meshes as GT videos for comparison.

- **MeadText** [90]. This dataset is extended from Mead [143] dataset by labeling the speaker emotional status and facial action unit (FAU) with natural language descriptions. MEAD [143] is a high-quality emotional talking-face dataset, including recorded videos of different actors speaking with 8 different emotions at 3 intensity levels. Note that the MeadText dataset is created according to manual template, for more natural and diverse descriptions users can leverage GPT-4V following the protocol of InstructAvatar [149].
- **RAVEDESS** [89]. There are a total of 24 professional actors (12 female, 12 male) covering over 1440 utterances in a neutral North American accent. 8 speech emotions includes calm, happy, sad, angry, fearful, surprise, disgust and neutral expressions are produced at two levels of emotional intensity (normal, strong). For convenience, we choose speech videos of the first 6 actors as the evaluation dataset.

Implementation Details. The videos are sampled at a rate of 25 FPS, and the audios are pre-processed to 16 kHz for all stages of our system. The training of the audio-visual instruction module is divided into two stages. In the first stage, the audios are fed to HuBERT [58] for speech feature extraction. Then, the Q-Former is pre-trained to contrastively align acoustic features to visual facial descriptions. Subsequently, the Q-Former is frozen, and we fine-tune the input projection layer of LLaMA-7b to achieve caption prediction. To initialize the LLaMA-7b model, we use Vicuna [23], an open-source text-based LLM widely utilized in dialogue generation. To enhance model performance, we leverage common text data augmentation techniques such as synonym replacement during the training stage.

For the talking face synthesis network, we adopt the model architecture of EMOTE [34] as our basic facial motion generation network. We adapt the framework with disentangled speech content space and content irrelevant space. For speech content extraction, we utilize the state-of-the-art pretrained ASR network Wav2Vec 2.0 [6] to extract the raw waveform and compress features with temporal convolutions following a similar protocol to EMOTE [34]. For speech

style extraction, we follow the architecture design of StyleTalk [91] and leverage the linear styling network from EMOTE [34] as a teacher network for knowledge distillation. Within the content irrelevant space, the training schedule of our contrastive instruction-style alignment and diffusion module is adapted from DALL-E2 [111]'s open-source implementation of diffusion prior. Specifically, the diffusion loss weight λ^{i2s} is set to 30 to balance optimization loss. Similar to the first stage, we also employ the same data augmentation approach to facilitate robust performance. As our focus in this work is on modeling speaking styles, the poses and speaker identity are set to a neutral state during both the training and inference stages. Both our models are implemented in PyTorch [105] and trained using 80G Tesla A100 GPUs. In our experiment, training the Audio-Visual Instruction network requires 12 hours, whereas training the instruction-following synthesis network takes 48 hours. Regarding inference time, processing a 30-second audio clip necessitates approximately 7.14 seconds for the Audio-Visual Instruction network to predict an instruction, and roughly 43.06 seconds for the synthesis network to generate the final video.

Comparison Methods. We compare our methods with state-of-the-art template-based models that support speech conditional 3D talking face generation, including MeshTalk [113], FaceFormer [41], CodeTalker [163], and EmoTalk [108].

MeshTalk [113] introduces a cross-modality disentanglement mechanism to generate realistic face animation. FaceFormer [41] devises a transformer-based architecture capable of synthesizing realistic 3D facial motions. CodeTalker[163] incorporates the codebook technique[39] to enhance the accuracy of lip movements. EmoTalk [108] employs an emotional disentanglement strategy using one-hot emotional labels for face animation. We also present a version of our approach that does not utilize a large language model. Instead, we directly employ the audio embedding obtained by Q-Former as the instruction source for the synthesis network, replacing its original language instruction input. For fair comparison, we utilize the audio embedding after contrastive audio-visual instruction alignment as a strong baseline. This alternative approach is referred to as AVI-Talking (w/o LLM).

5.3.2 Quantitative Evaluation

Evaluation Metric. We validate our method from the perspectives of both instruction generation capability and talking face synthesis quality.

- **Audio-Visual Instruction Prediction.** Metrics that have popularly been involved in the field of natural language generation (NLG) task are chosen to evaluate our method. Specifically, we include $BLEU_1$, $BLEU_4$ [104], $METEOR$ [7], $ROUGE_l$ [82], $CIDEr$ [147] and $SPICE$ [2]. The calculation procedure of these metrics is included in Sec. 2.2.5.

- **3D Talking Face Synthesis.** To assess visual fidelity, we utilize standard GAN metrics: **FID** [53] and **KID** [112] on face regions of rendered images following the evaluation protocol of previous work [4]. Additionally, to evaluate generation diversity, we report **Diversity** scores [3], measuring the extent of expression diversity generated for a given clip of human speech. Specifically, distances across predicted style features for the same audio with different noises are calculated. Moreover, we adopt **LSE-D** [110] to evaluate lip synchronization performance. The calculation procedure of **LSE-D** is illustrated in Sec. 2.3.1. Following the EmoTalk [108], we also calculate the vertex error to validate the expressiveness of our results. Due to their different mesh topology involved in Emotalk, we manually align their mesh with FLAME in canonical space and apply some shape transfer techniques [83] to compare with other approaches.

  Here we briefly introduce the calculation of **KID** and **FID**. Both FID (Fréchet Inception Distance) and KID (Kernel Inception Distance) serve as metrics to gauge the likeness between two collections of images. FID employs activation features from a pre-trained Inception-v3 network to calculate the Fréchet distance between their respective sets of feature vectors. Conversely, KID utilizes the Maximum Mean Discrepancy (MMD) metric, a measure of the distance between probability distributions, to assess the divergence between real and generated image distributions within feature space. In both cases, lower scores signify superior performance.

Evaluation Results. Regarding the synthesis of talking faces, our study reports quantitative results for MeadText [90] and RAVEDESS [89] in Table 5.1. Notably,

Table 5.1: The quantitative results on MeadText [90] and RAVEDESS [89].

Method	MeadText [90]				RAVEDESS [89]			
	FID ↓	KID ↓	LSE-D ↓	EVE ↓ ($\times 10^{-6}$)	FID ↓	KID ↓	LSE-D ↓	EVE ↓ ($\times 10^{-6}$)
MeshTalk[113]	201.06	0.3601	10.51	-	134.47	0.2831	9.19	-
EmoTalk[108]	124.41	0.2118	**8.37**	3.129	122.95	0.1929	**8.51**	3.337
CodeTalker[163]	68.68	0.0658	<u>8.38</u>	5.211	<u>46.90</u>	<u>0.0711</u>	8.99	4.059
FaceFormer[41]	<u>68.35</u>	<u>0.0611</u>	9.08	5.234	47.78	0.0721	8.85	4.036
GT	-	-	9.36	-	-	-	9.05	-
w/o LLM	12.91	0.0205	8.95	-	16.59	0.0259	<u>8.56</u>	-
AVI-Talking	**12.53**	**0.0190**	9.06	**2.228**	**15.94**	**0.0225**	8.81	**2.323**

our method demonstrates outstanding performance across most metrics on both datasets. However, our approach may exhibit comparatively weaker lip-sync performance, particularly in terms of LSE-D, when compared to other methods. We attribute this discrepancy partly to the strong preference bias for neutral expressions in SyncNet [110], which is pre-trained on predominantly expressionless videos. Unlike these methods, our synthesis results encompass expressive facial details, potentially leading to lower scores. Furthermore, our approach achieves LSE-D scores close to those of ground truth videos on both datasets, suggesting robust generation of precise lip-sync videos. It's worth noting that our full model outperforms the variant that removes LLMs, underscoring the effectiveness of incorporating LLMs as an additional audio-visual instruction agent in our system.

5.3.3 Qualitative Evaluation

Qualitative Analysis. Subjective evaluation is crucial for validating model performance in generative tasks. We encourage readers to refer to our supplementary materials for additional demo videos and comparison results. In Figure 5.5, we present comparison results of our method against previous state-of-the-art approaches in three cases. In the top row are ground truth videos. Our generated

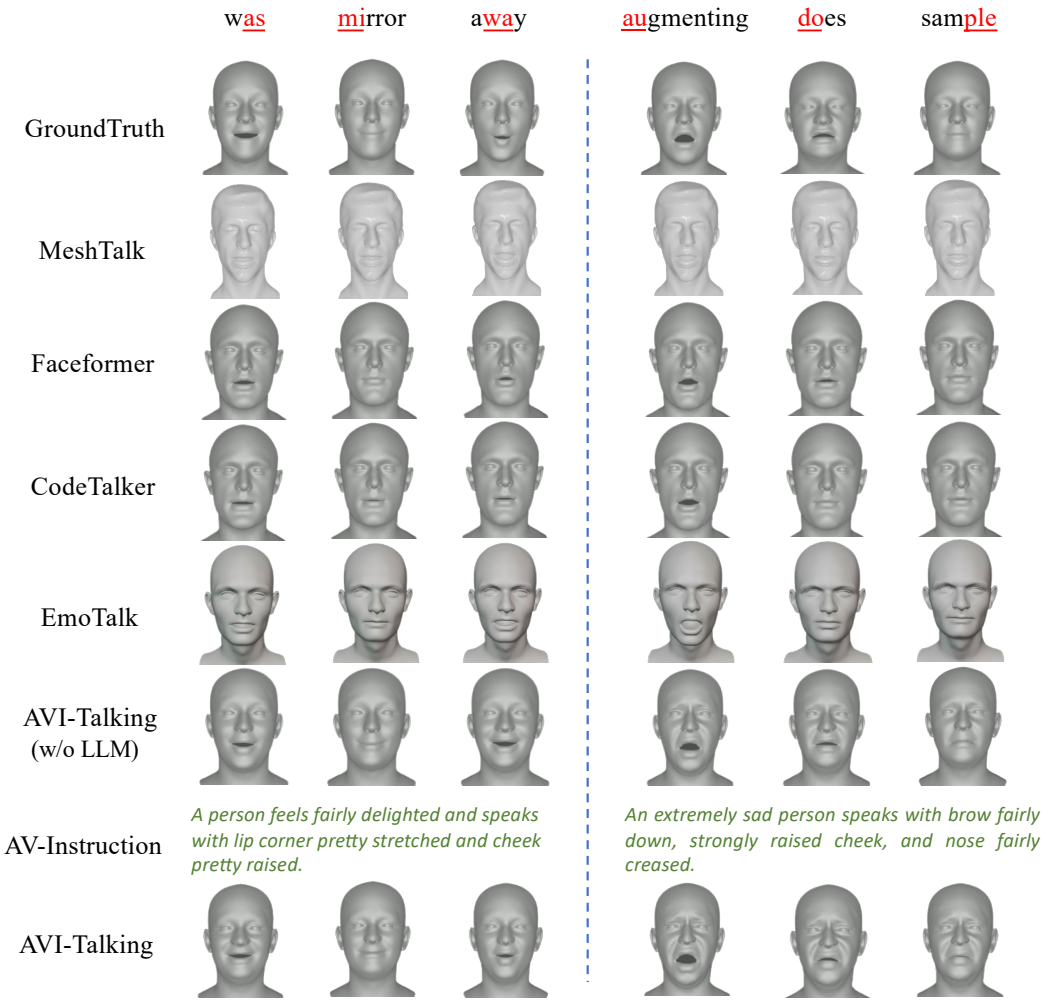

Figure 5.5: Qualitative Results.

audio-visual instructions are shown in green line. In the bottom row demonstrates synthesis results guided by above instructions. Compared to other competitive approaches, our method achieves superior detailed expressions. Notably, our system is capable of generate facial movements distinct from Ground Truth but convey consistent speaking state (See second and third case). It can be seen that our method produces plausible audio-visual instructions and generates expressive facial details aligned with the speaker's state. Regarding lip synchronization performance, we observe that CodeTalker [163] or Faceformer [41] may generate more natural pronunciation in expressionless states. However, when involving

Table 5.2: User study measured by Mean Opinion Scores. Larger is higher, with the maximum value to be 5.

MOS on \ Approach	MeshTalk [113]	EmoTalk [108]	CodeTalker [163]	**Ours**
Lip Sync Quality	2.43	2.83	3.13	**3.23**
Movement Expressiveness	2.83	3.0	2.53	**3.27**
Expression Consistency	2.37	3.03	2.33	**3.50**

emotional states, slight distortions in lip movements can be observed (e.g., the stretching of lip corners during happy emotions). This observation aligns with the LSE-D scores in the quantitative evaluation presented in Table 5.1. Nevertheless, our approach still achieves competitive synthesis results compared to others and approaches the performance of ground truth videos, thus validating the effectiveness of our approach in lip synchronization. When compared with our variant version (without LLM), the synthesis results exhibit richer facial details, such as raised cheeks and creased noses (as observed in the happy and sad cases). We believe that this phenomenon may arise from the complexity of information embedded within human speech. While this complexity may slightly compromise the performance of the synthesis network, resulting in the loss of some subtle details.

User Study. We conducted a user study involving 15 participants to gather their opinions on 30 videos generated by our method alongside three competing methods. Among these, twenty videos were created using randomly selected speaker audios from the test set of MeadText, while the remaining ten were sourced from RAVEDESS. We utilized the well-established Mean Opinion Scores (MOS) rating protocol. Participants were tasked with providing ratings on a scale of 1 to 5 for three specific aspects of each video: (1) Lip Sync Quality, (2) Movement Expressiveness, and (3) Expression Consistency. Lip sync quality evaluates mouth movements in sync with speech content, movement expressiveness assesses facial detail richness, and expression consistency measures the alignment between facial movements and speaker speech expressions.

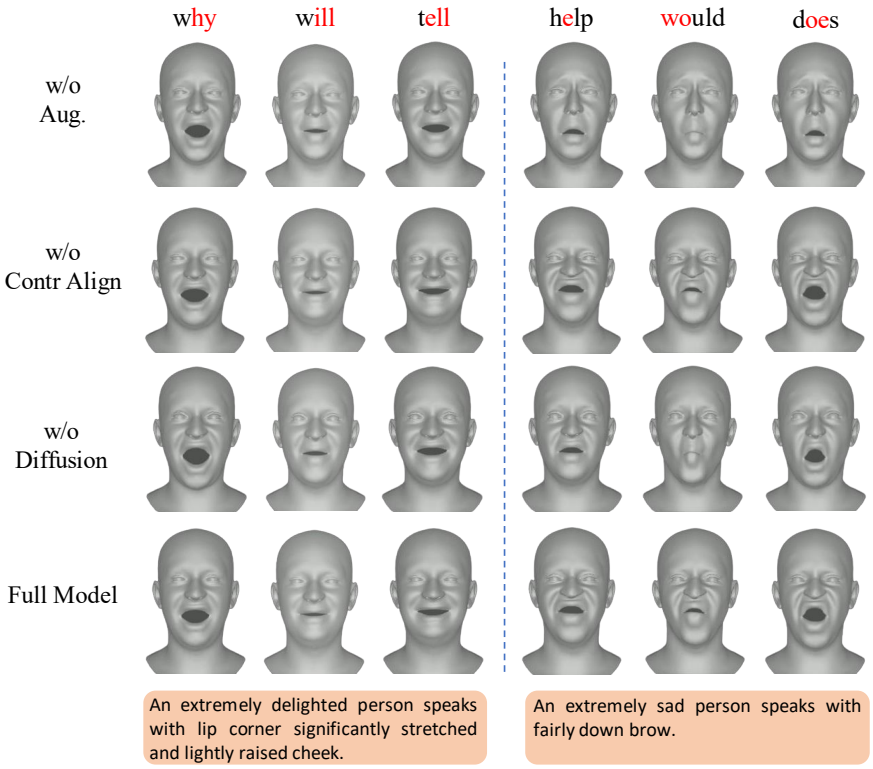

Figure 5.6: Ablation Study.

The results are presented in Table 5.2. MeshTalk [113] scores the lowest across all aspects, possibly attributed to the architecture design of its naive UNet. By incorporating transformer blocks, EmoTalk [108] and CodeTalker [163] achieve higher lip-sync scores. Regarding movement expressiveness and expression consistency, our model significantly surpasses other approaches, owing to its carefully derived audio-visual instruction strategy. Overall, our AVI-Talking model outperforms its counterparts in expressive synthesis, highlighting the effectiveness of our approach.

5.3.4 Further Analysis

Ablation Study We conduct ablation studies on both stages of our system, wherein we systematically remove three crucial components from each stage to evaluate the effectiveness of our framework design.

Table 5.3: Ablation over model design of Audio-Visual Instruction stage.

Metric	w/o Aug.	w/o LLaMA	w/o Q-Former	Full
$BLEU_1 \uparrow$	45.4	32.4	41.4	**47.4**
$BLEU_4 \uparrow$	**12.7**	7.1	10.4	11.4
$METEOR \uparrow$	21.5	16.1	20.7	**22.0**
$ROUGE_l \uparrow$	38.0	28.0	36.0	**38.4**
$CIDEr \uparrow$	**54.5**	32.8	53.0	49.3
$SPICE \uparrow$	34.8	27.4	36.4	**40.9**

Table 5.4: Ablation over model design of Talking Face Synthesis stage.

Method	LSE-D ↓	Diversity ↑	FID ↓	KID ↓
w/o Aug.	9.11	0.433	14.18	0.0192
w/o Diffusion	9.21	0	18.72	0.0268
w/o Cont Align	9.07	0.373	13.37	0.0190
Full (Ours)	**9.06**	**0.435**	**12.53**	**0.0190**

Audio-Visual Instruction Module. We conduct experiments on the first stage model (1) w/o text augmentation; (2) w/o LLaMA generator and (3) w/o Q-Former alignment. For the setting without the LLaMA base model, we adopt the BLIP2 training paradigm [78] and utilize image-grounded text generation loss for instruction generation. The numerical results on the MeadText dataset [90] are presented in Table 5.3. We find that without text data augmentation, the model tends to overfit to a sub-optimal point, leading to slightly worse performance. But the overall impact for the performance is marginal considering that metrics of $BLEU_4$ and $CIDEr$ are even better. Removing the LLaMA model results in the loss of rich contextual knowledge, thereby also causing inferior performance. Furthermore, without the Q-Former contrastive alignment strategy, the extraction and alignment of speech features to text embedding become inadequate, introducing significant training difficulties and resulting in significantly inferior performance.

3D Talking Face Synthesis. For the second stage, we train and evaluate the talking face synthesis network by removing (1) text augmentation, (2) the diffusion prior network, and (3) contrastive alignment. The numerical results on the MeadText dataset [90] are demonstrated in Table 5.4, and visualization results are depicted in Figure 5.6. Similar to the first stage, without text data augmentation, the synthesis results suffer from inferior performance on all metrics. Visualization results in the first row illustrate that the absence of augmentation tends to inadequately capture the smiling lip corner motion (See the first case in the left column). Without employing the diffusion strategy, the generation process becomes deterministic, leading to a lack of diversity. We also observe significantly reduced performance on other metrics, possibly due to the diverse generation nature of this problem. Visualizations in Figure 5.6 indicate that without adopting the diffusion strategy, the network tends to produce conservative generations, where the lip corner is not as well stretched as in our full model (See the first case in the left column). Removing contrastive alignment also results in inferior outcomes, highlighting its effectiveness in boosting generation performance.

Visualization of Aligned Speech Features. To further analyze the performance of Audio-Visual Instruction design, we visualize the intermediate speech features that are contrastively aligned using Q-Former. In particular, as discussed in Sec. 5.2.3, the contrastive audio-visual instruction alignment aims to extract audio embeddings closely relevant to the visual instructions. Consequently, the resulting audio embeddings are expected to include rich speaker state information.

As shown in Figure 5.7, the audio samples are from male utterances in the MeadText dataset, focusing on five typical speaking emotions. Notably, the aligned speech features corresponding to each specific emotion exhibit closely clustered patterns. We present samples of utterances representing five typical emotions. Notably, there is a discernible clustering pattern observed among embeddings associated with the same emotional type. It is interesting that speech features belonging to the happiness class exhibit particularly close clustering, which could be attributed to the distinct characteristics of a happy voice.

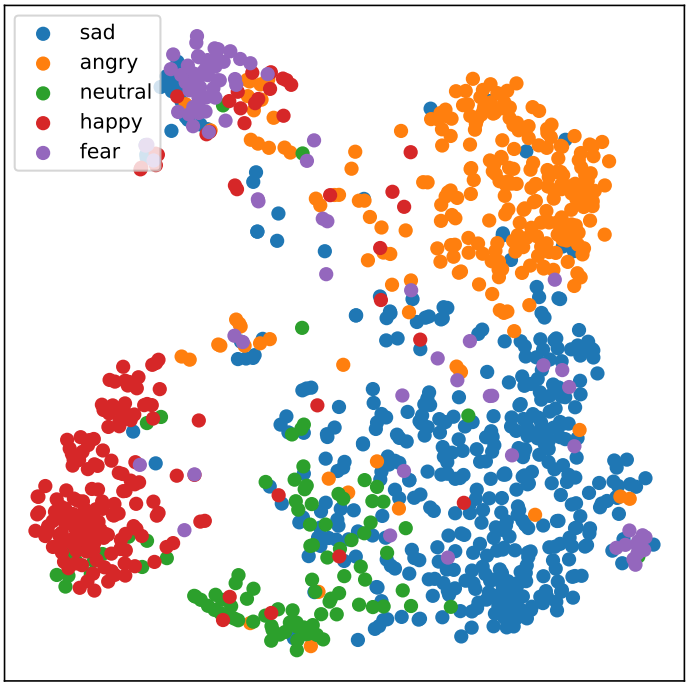

Figure 5.7: Visualizations of t-SNE embeddings derived from aligned speech features using Q-Former.

Generation Diversity of Talking Face Instruction System. In Table 5.4, we illustrate the pivotal role of diffusion strategy in enhancing generation diversity. Additionally, in Figure 5.8, the bottom row showcases the audio-visual instruction, while the rows above demonstrate generation variations using the same text instruction. The left columns display the sad speaker status, where different lip curves are predicted, while the right columns demonstrate an angry case with varying eyebrow and cheek movements. We present visualizations showcasing diverse synthesis. Observing the left column, it shows that multiple lip curves can be synthesized for instructions conveying disappointing emotions. Similarly, the right column demonstrates varied eyebrow and cheek movements in response to text instructions suggesting anger. These outcomes validate the capability of the talking face synthesis system to produce diverse results.

OOD Analysis of Talking Face Instruction System. To further assess the

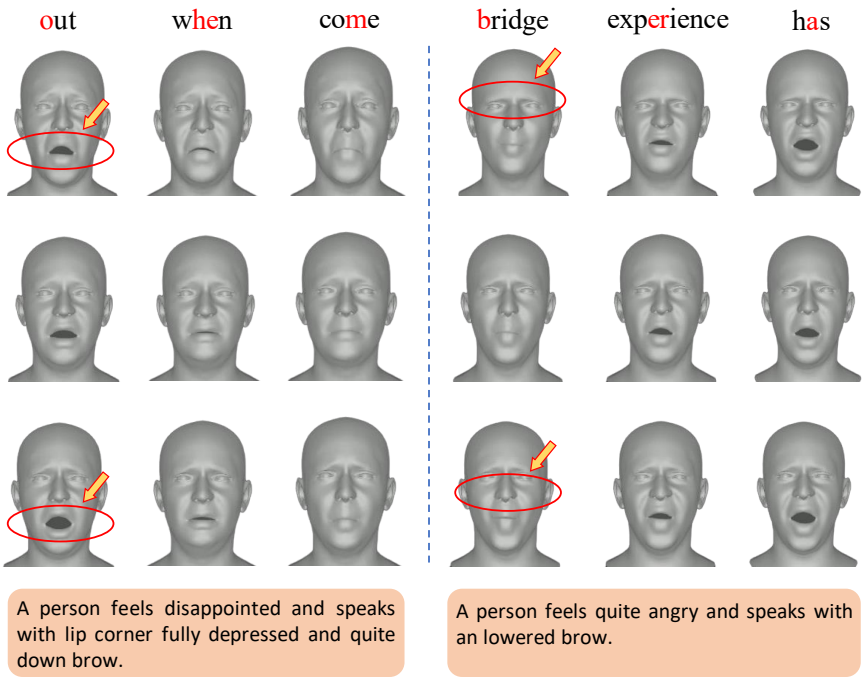

Figure 5.8: Diverse generation results of the talking face instruction system are depicted.

generalizability of our proposed talking face synthesis module, we conducted experiments with out-of-distribution (OOD) instructions. Unlike the instructions in our dataset, which explicitly describe facial movements, we also explored visual instructions indicated by abstract concepts. As shown in Figure 5.9, within each row, we present instructed synthesis outcomes for the same speaker's speech, encompassing four distinct out-of-distribution instructions. The initial three rows showcase various successful cases while the final row illustrates an instance where the model misinterprets the instruction. Our model demonstrates the ability to capture the implicit speaking state of the speaker in the first three rows, yielding plausible synthesis results. This success can be attributed to the adoption of the diffusion mechanism and the structural similarity of natural language embeddings. However, when faced with particularly complex and abstract instructions, our model tends to misinterpret the implied speaking states as seen in the last row.

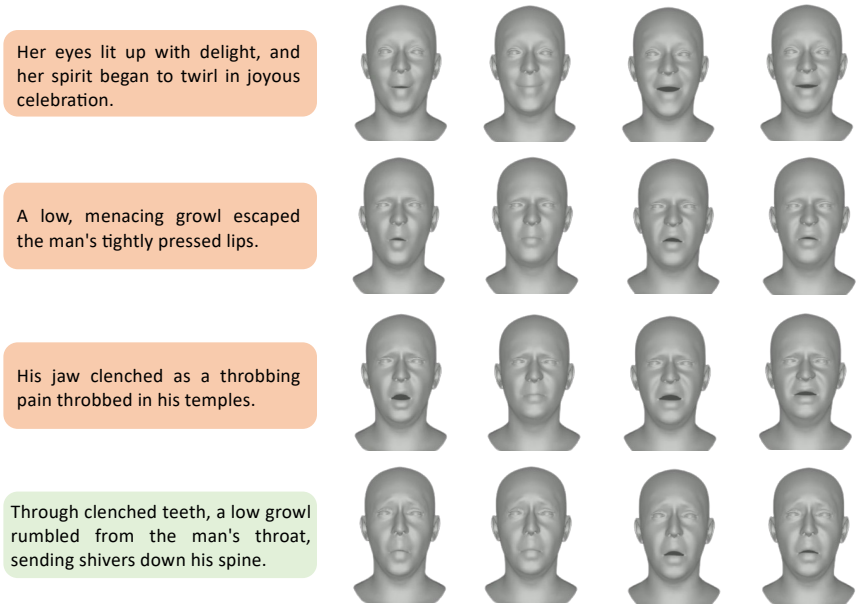

Figure 5.9: Visualization of Out-of-Distribution (OOD) results from the Talking Face Instruction System.

5.4 Related Work

While methods such as [41, 163, 31, 108] successfully synchronize facial motion with the driven audio, they often learn deterministic models, resulting in rigid motion within speech-irrelevant regions, leading to unnatural synthesis. To address these limitations, recent approaches [129] introduce a diffusion mechanism for its remarkable generative capability, yielding diverse high-quality synthesis results [135, 76].

However, while modeling various poses and expressions, these approaches neglect to capture the emotional content implied within the audio. Furthermore, methods relying on end-to-end diffusion, from reference video or style embedding to parametric models, lack explainability. Therefore, we propose integrating a large language model into our system to firstly generate an interpretable audio-visual instruction, which is leveraged to guide the speech-driven 3D talking head generation. To augment emotion awareness, we apply the diffusion process coupled with contrastive learning solely to the speech-irrelevant space.

Recent advancements in talking face generation have demonstrated the lan-

guage model's capacity to generate multi-modal content [179] and synthesize facial motions [99, 141]. Typically, these approaches involve deriving special tokens for another modality and learning a projection layer to align them with language space [99]. However, this process demands substantial paired data and intricate training techniques for effective alignment. In contrast to these methodologies, our approach takes a direct path by predicting the text description of emotional status and facial details. This eliminates the need for challenging cross-modal alignment procedures, which also provides enhanced explainability and flexibility to users.

For convenience, we have provided a list of related works on AVI-Talking in Table 5.5. Our model distinguishes itself as the only 3D speech-driven talking face system that comprehensively supports emotion, language instruction, and explainability within a single, holistic framework.

	Year	With Emotion	3D	Text Control	Explainability
Mead [144]	2020	✓	✗	✗	✗
GC-AVT [80]	2022	✓	✗	✗	✗
EAMM [65]	2022	✓	✗	✗	✗
Sinha *et al.* [121]	2022	✓	✗	✗	✗
TalkClip [90]	2023	✓	✗	✓	✗
MeshTalk [113]	2021	✗	✓	✗	✗
EmoTalk [108]	2023	✓	✓	✗	✗
CodeTalker [163]	2023	✗	✓	✗	✗
FaceFormer [41]	2022	✗	✓	✗	✗
AVI-Talking	2024	✓	✓	✓	✓

Table 5.5: Relation of AVI-Talking with State-of-the-Art Methods

5.5 Summary

In this paper, we propose **AVI-Talking**, an **Audio-Visual Instruction** system for expressive 3D **Talking** face generation. In contrast to the aforementioned approaches, we explore harnessing the generative power of large language models

(LLMs) to act as a multi-modality reasoning engine. This will actively hallucinates diverse and plausible facial details based on the emotional content of the input audio, thereby offering a more comprehensive and nuanced synthesis. We emphasize several appealing properties of our framework: 1) We address the speech-driven expressive talking face generation by introducing an intermediate visual instruction, which decomposes the challenging audio-to-visual generation into two stages with clear learning objective. 2) A soft prompting strategy is derived to harness the prior contextual knowledge underlying LLMs for speaker talking state comprehension. 3) The disentangled talking prior learning procedure ensures complementary integration of lip-sync movements and audio-visual instruction. 4) A diffusion prior network is introduced to map audio-visual instructions to latent distribution of content irrelevant space.

Future Work. In this paper, we have investigated into specifying a pretrained Large Language Model (LLM) for cross-modal audio-visual generation using finetuning techniques. Recent studies [77] higlight the remarkable capability of Retrieval Augmented Generation (RAG) in injecting knowledge into Large Language Models (LLMs). Future research will involve comparing the effectiveness of RAG and fine-tuning performance, particularly tailored for this task.

Meanwhile, recent works [5] suggest that visual foundation models can yield competitive results, provided a robust visual tokenizer is utilized. Consequently, future research will delve into directly tokenizing stylized embeddings within the content-irrelevant space and fine-tuning general visual foundation models for expressive talking face synthesis. In this way, the model might be able to circumvent relying on specific audio-visual instruction dataset, thereby achieving superior performance with high generality.

6 Conclusion

This paper targets realistic video generation by exploiting the implicit information of human speech. We propose to leverage explicit visual cues to facilitate audio-visual association learning through a framework with three key components. The overall architecture follows an encoder-decoder structure with a disentangled bottleneck latent space, and explicit visual cues are utilized to facilitate audio-visual association learning via contrastive learning.

According to this framework, two networks are carefully adapted for 2D and 3D talking face synthesis, taking into account speaker identity and talking status. The Speech2Talkingface is a generative framework capable of not only inferring appearances but also driving faces to talk with speech. It synchronize speech with visual representations through two branches and uses a style-based generator for balancing multi-modal information. Two latent spaces, the identity space and the identity-irrelevant space, are defined through the reconstruction training of visual features. Contrastive loss is employed for modality synchronization within the these spaces.

AVI-Talking is an Audio-Visual Instruction System for expressive Talking face generation. It bridge the audio-visual modality gap with an intermediate visual instruction representation. The framework decomposes the audio-to-video generation into two stages, each with a distinctive objective. By presenting visual instruction as an intermediate output, our system not only enhances model interpretability but also grants users the flexibility to specify desired instructions or modifications, enriching user interaction and greater customization. A contrastive instruction-style alignment and diffusion strategy further refines the joint representation towards the distribution of the pre-trained talking prior.

Overall, the first network effectively integrates the identity information of the human voice, while the second network extracts the underlying talking status information from the speaker's speech. The design of both networks follows our proposed framework, validating its effectiveness for realistic facial video synthesis.

The Speech2Talkingface network employs data augmentation techniques to identify the speech content space, while AVI-Talking directly relies on the pre-trained Wav2Vec-2.0 for content relevant information extraction. Using the pretrained Wav2Vec-2.0 network provides a well speech-related prior, easing the learning process of content-relevant space formation. Additionally, Speech2Talkingface uses a convolution-based architecture suitable for 2D image synthesis, while AVI-Talking leverages a transformer-based network for dynamic vertex modeling. To bridge the modality gap, Speech2Talkingface fine-tunes the entire system, while AVI-Talking uses a diffusion-based network, demonstrating the effectiveness of diffusion models in closing the modality gap.

6.0.1 Outlook

1. **Real-time Talking Face Applications.** In current framework, the computation of Audio-Visual Instruction module suffers from heavy overload. Deploying such system to real-time applications will lead to prohibitively calculation burden. Therefore, it is necessary to devise some strategies to improve the running efficiency of this module in future work.

2. **Unified Framework.** In our AVI-Talking work, we utilize separate speech encoders: HuBERT within the audio-visual instruction module and Wav2Vec-2.0 within the talking face synthesis module. However, is it possible to achieve this functionality with a unified speech extractor? Future research will explore the potential of using a single network to simultaneously learn both talking states and acoustic content.

3. **Facilitating Audio-to-Visual Instruction.** This work leverages publicly available language generator LLama [136] as the backbone to synthesize visual instructions. To promote speech information comprehension, our work compose an audio encoder to prompt this language model. However, the devised module is just a preliminary exploration. Recent developed multi-modal language models such as LLama3 might be an option to achieve better performance.

4. **Image Quality Enhancement with Diffusion Transformers.** Recent works in generative models have demonstrated the capability of Diffusion

Transformers (DiTs) [107] with the help of various newly design training techniques [190, 46, 18, 19] such as flowing matching or positioning encoding. With remarkable scalability, these architectures support high-quality video synthesis [180, 102]. It is worth exploring leveraging these backbones in Speech2Talking-Face work to improve the resolution and generation quality in our tasks.

5. **Multi-modal Talking Face Instruction System**. Future work will exploit the mutual benefits of human voice and visual cues to develop a multi-modal talking face instruction system, addressing the ambiguity of vocal features with the explicit semantic meaning of visual cues.

6. **Long-range Sequential Modeling**. In the Speech2Talkingface application, the model currently determines speaker identity from voice clips of less than 3 seconds. Future work will attempt to capture speaker identity with long-range sequential modeling to utilize more relevant information.

7. **Addressing Human Race Bias**. The unbalanced distribution of speakers in the Speech2Talkingface framework may introduce human race bias, leading to ethical issues. Future research will explore larger datasets for objective distribution and safeguard the pre-trained model for misuse.

8. **Incorporating Head Motion**. The AVI-Talking application assumes static head poses, which is not always accurate. To synthesize more vivid results, future work will include head motion descriptions within the audio-visual instructions to incorporate talking head dynamics.

9. **Facial Appearance Modeling**. Unlike Speech2Talkingface, the AVI-Talking framework does not account for facial appearance. Future work will consider integrating a neural renderer for facial appearance modeling.

10. **Speech-to-Gesture Modeling**. Facial movements are only one part of the broader spectrum of human communication. Gestures, including arm movements and hand actions, are closely linked to speech and serve as powerful expressive tools in human interaction. Therefore, it is crucial to incorporate these elements into virtual digital human systems to create more natural and effective communication. Moreover, recent advancements in diffusion

models and large language models (LLMs) offer innovative techniques to enhance the development of these systems.

Bibliography

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[3] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. *arXiv preprint arXiv:2312.08459*, 2023.

[4] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21263–21273, 2024.

[5] Anonymous. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gzqrANCF4g`.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[7] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[8] Çağdaş Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulović. A framework for the robust evaluation of sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65. IEEE, 2020.

[9] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[10] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016.

[11] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018.

[12] Elodie F Briefer. Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology*, 288(1):1–20, 2012.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[14] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *CVPR*, 2020.

[15] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24 (4):1283–1302, 2005.

[16] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, 2023.

[17] Soham Chattopadhyay, Arijit Dey, Hritam Basak, et al. Optimizing speech emotion recognition using manta-ray based feature selection. *arXiv preprint arXiv:2009.08909*, 2020.

[18] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[19] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-δ: Fast and controllable image generation with latent consistency models, 2024.

[20] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018.

[21] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019.

[22] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020.

[23] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2 (3):6, 2023.

[24] Hyeong-Seok Choi, Changdae Park, and Kyogu Lee. From inference to generation: End-to-end fully self-supervised generation of human face from speech. *ICLR*, 2020.

[25] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[26] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[27] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[28] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV Workshop*, 2016.

[29] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017.

[30] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *INTERSPEECH*, 2018.

[31] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019.

[32] Prajakta P. Dahake, Kailash Shaw, et al. Speaker dependent speech emotion recognition using mfcc and support vector machine. In *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pages 1080–1084, 2016. doi: 10.1109/ICACDOT.2016.7877753.

[33] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

[34] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–13, 2023.

[35] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[36] Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i Nieto. Wav2pix: Speech-conditioned face generation using generative adversarial networks. In *ICASSP*, 2019.

[37] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. End-to-end generation of talking faces from noisy speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1948–1952. IEEE, 2020.

[38] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 24:3480–3490, 2021.

[39] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

[40] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015.

[41] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022.

[42] Gunnar Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Number 2. Walter de Gruyter, 1971.

[43] Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, S Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 18225–18250. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/3a7f9e485845dac27423375c934cb4db-Paper-Conference.pdf`.

[44] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

[45] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645, 2023.

[46] Peng Gao, Le Zhuo, Chris Liu, , Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

[47] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021.

[48] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[49] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable

expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20914–20923, 2023.

[50] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[51] Jiaqi Hao, Shiguang Liu, and Qing Xu. Controlling eye blink for talking face generation via eye conversion. In *SIGGRAPH Asia 2021 Technical Communications*, pages 1–4. 2021.

[52] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[53] 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. *Advances in neural information processing systems*, 30, 2017.

[54] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851, 2020.

[55] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[56] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[57] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. Face-voice matching using cross-modal embeddings. In *ACMMM*, 2018.

[58] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[59] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[60] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023.

[61] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[62] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127:1767–1779, 2019.

[63] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[64] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *CVPR*, 2021.

[65] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[66] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. Putting the face to the voice': Matching identity across modality. *Current Biology*, 13(19):1709–1714, 2003.

[67] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[68] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[69] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

[70] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7: 117327–117345, 2019.

[71] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *ACCV*, 2018.

[72] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *TOG*, 2018.

[73] Andrew Koh, Xue Fuzhao, and Chng Eng Siong. Automated audio captioning using transfer learning and reconstruction latent space similarity regularization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7722–7726. IEEE, 2022.

[74] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[75] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1428–1436, 2019.

[76] Max W. Y. Lam, Jun Wang, Dan Su, and Dong Yu. BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=L7wzpQttNO`.

[77] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[78] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[79] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL `https://doi.org/10.1145/3130800.3130813`.

[80] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022.

[81] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multimodal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[82] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[83] Jiangke Lin, Yi Yuan, and Zhengxia Zou. Meingame: Create a game character face from a single portrait. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 311–319, 2021.

[84] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[85] Shiguang Liu. Audio-driven talking face generation: A review. *Journal of the Audio Engineering Society*, 71(7/8):408–419, 2023.

[86] Shiguang Liu and Jiaqi Hao. Generating talking face with controllable eye movements by disentangled blinking feature. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[87] Shiguang Liu and Huixin Wang. Talking face generation via facial anatomy. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–19, 2023.

[88] Zhen-Tao Liu, Qiao Xie, Min Wu, Wei-Hua Cao, Ying Mei, and Jun-Wei Mao. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing*, 309:145–156, 2018.

[89] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13 (5):e0196391, 2018.

[90] Yifeng Ma, Suzhen Wang, Yu Ding, Bowen Ma, Tangjie Lv, Changjie Fan, Zhipeng Hu, Zhidong Deng, and Xin Yu. Talkclip: Talking head generation with text-guided expressive speaking styles. *arXiv preprint arXiv:2304.00334*, 2023.

[91] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv preprint arXiv:2301.01081*, 2023.

[92] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023.

[93] Lauren W Mavica and Elan Barenholtz. Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2):307, 2013.

[94] Larry Medsker and Lakhmi C Jain. *Recurrent neural networks: design and applications*. CRC press, 1999.

[95] Givi Meishvili, Simon Jenni, and Paolo Favaro. Learning to have an ear for face super-resolution. In *CVPR*, 2020.

[96] Omar Mohamed and Salah A Aly. Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset. *arXiv preprint arXiv:2110.04425*, 2021.

[97] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *ECCV*, 2018.

[98] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, 2018.

[99] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10083–10093, 2023.

[100] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *CVPR*, 2019.

[101] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[102] OpenAI. Video generation models as world simulators, 2024.

[103] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human

feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[104] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[105] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[106] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[107] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[108] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023.

[109] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 80–88, 2017.

[110] 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.

[111] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[112] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[113] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021.

[114] Najmeh Sadoughi and Carlos Busso. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing*, 12(4):1031–1044, 2019.

[115] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[116] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[117] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[118] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[119] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[120] Sanjana Sinha, Sandika Biswas, and Brojeshwar Bhowmick. Identity-preserving realistic talking face generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2020.

[121] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *arXiv preprint arXiv:2205.01155*, 2022.

[122] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018.

[123] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 919–925, 2019.

[124] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[125] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[126] Yasheng Sun, Hang Zhou, Ziwei Liu, and Hideki Koike. Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation. In *IJCAI*, volume 2, page 4, 2021.

[127] Yasheng Sun, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, and Hideki Koike. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *arXiv preprint arXiv:2308.00906*, 2023.

[128] Yasheng Sun, Wenqing Chu, Hang Zhou, Kaisiyuan Wang, and Hideki Koike. Avi-talking: Learning audio-visual instructions for expressive 3d talking face generation. *IEEE Access*, 12:57288–57301, 2024. doi: 10.1109/ACCESS.2024.3390182.

[129] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Gaetan Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic

3d facial animation and head pose generation via diffusion models. *arXiv preprint arXiv:2310.00434*, 2023.

[130] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL `https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html`.

[131] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[132] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22146–22156, 2023.

[133] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[134] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017.

[135] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[136] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation

language models. *ArXiv*, abs/2302.13971, 2023. URL `https://api.semanticscholar.org/CorpusID:257219404`.

[137] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[138] 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.

[139] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *IJCV*, 2019.

[140] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128:1398–1413, 2020.

[141] Duomin Wang, Bin Dai, Yu Deng, and Baoyuan Wang. Agentavatar: Disentangling planning, driving and rendering for photorealistic avatar agents. *arXiv preprint arXiv:2311.17465*, 2023.

[142] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17979–17989, 2023.

[143] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

[144] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020.

[145] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.

[146] Lijuan Wang, Xiaojun Qian, Wei Han, and Frank K Soong. Synthesizing photo-real talking head via trajectory-guided sample selection. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[147] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4195–4203, 2019.

[148] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[149] Yuchi Wang, Junliang Guo, Jianhong Bai, Runyi Yu, Tianyu He, Xu Tan, Xu Sun, and Jiang Bian. Instructavatar: Text-guided emotion and motion control for avatar generation. *arXiv preprint arXiv:2405.15758*, 2024.

[150] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[151] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[152] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. Disjoint mapping network for cross-modal matching of voices and faces. *ICLR*, 2019.

[153] Yandong Wen, Bhiksha Raj, and Rita Singh. Face reconstruction from voice using generative adversarial networks. In *NeurIPS*, 2019.

[154] John R Westbury, Michiko Hashi, and Mary J Lindstrom. Differences among speakers in lingual articulation for american english. *Speech Communication*, 26(3):203–226, 1998.

[155] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018.

[156] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[157] Cho-Ying Wu, Chin-Cheng Hsu, and Ulrich Neumann. Cross-modal perceptionist: Can face geometry be gleaned from voices? In *CVPR*, 2022.

[158] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1478–1486, 2021.

[159] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2023.

[160] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[161] Lei Xie and Zhi-Qiang Liu. Realistic mouth-synching for speech-driven talking face using articulatory modelling. *IEEE Transactions on Multimedia*, 9 (3):500–510, 2007.

[162] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[163] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[164] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6619, 2023.

[165] Xuenan Xu, Mengyue Wu, and Kai Yu. A comprehensive survey of automated audio captioning. *arXiv preprint arXiv:2205.05357*, 2022.

[166] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331, 2024.

[167] Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip JB Jackson, and Mark D Plumbley. Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1230–1241, 2017.

[168] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023.

[169] Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7645–7655, 2023.

[170] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019.

[171] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan

Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-Aw0rrrPUF.

[172] Dan Zeng, Han Liu, Hui Lin, and Shiming Ge. Talking face generation with expression-tailored generative adversarial network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1716–1724, 2020.

[173] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[174] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022. URL https://api.semanticscholar.org/CorpusID:248496292.

[175] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023.

[176] 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*, pages 3661–3670, 2021.

[177] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[178] Rui Zhen, Wenchao Song, Qiang He, Juan Cao, Lei Shi, and Jia Luo.

Human-computer interaction system: A survey of talking-head generation. *Electronics*, 12(1):218, 2023.

[179] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens, 2023.

[180] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL `https://github.com/hpcaitech/Open-Sora`.

[181] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019.

[182] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *ICCV*, 2019.

[183] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021.

[184] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021.

[185] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018.

[186] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking head animation. *SIGGRAPH ASIA*, 2020.

[187] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[188] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. *arXiv preprint arXiv:1812.06589*, 2018.

[189] Hao Zhu, Aihua Zheng, Huaibo Huang, and Ran He. High-resolution talking face generation via mutual information approximation. *IJCAI*, 2020.

[190] Le Zhuo, Ruoyi Du, Xiao Han, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024.

Publication List

Refereed Journals

1. **Yasheng Sun**, Wenqing Chu, Hang Zhou, Kaisiyuan Wang and Hideki Koike. AVI-Talking: Learning Audio-Visual Instructions for Expressive 3D Talking Face Generation. IEEE Access, pages 57288-57301, April 2024.

Refereed International Conference Papers

2. **Yasheng Sun**, Hang Zhou, Ziwei Liu, and Hideki Koike. Speech2Talking-Face: Inferring and Driving a Face with Synchronized Audio-Visual Representation. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI '21), pages 1018–1024, August 2021.

Co-authored Publications

3. **Yasheng Sun**, Qianyi Wu, Hang Zhou, Kaisiyuan Wang, Tianshu Hu, Chen-Chieh Liao, Shio Miyafuji, Ziwei Liu, and Hideki Koike. Make Your Brief Stroke Real and Stereoscopic: 3D-Aware Simplified Sketch to Portrait Generation. In Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23), pages 388–396, October 2023.

4. **Yasheng Sun**, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. Masked Lip-Sync Prediction by Audio-Visual Contextual Exploitation in Transformers. In SIGGRAPH Asia 2022 Conference Papers (SA '22), pages 1–9, November 2022.

5. **Yasheng Sun**, Jiangke Lin, Hang Zhou, Zhiliang Xu, Dongliang He, Hideki Koike. ReEnFP: Detail-Preserving Face Reconstruction by Encoding Facial Priors. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV '23), pages 6118-6128, January 2023.

6. **Yasheng Sun**, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike. ImageBrush: Learning Visual In-Context Instructions for Exemplar-Based Image Manipulation. In Advances in Neural Information Processing Systems (NeurIPS '23), December 2023.

7. Hang Zhou, **Yasheng Sun**, Wayne Wu, Chen Change Loy, Xiaogang Wang, Ziwei Liu. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '21), pages 4176-4186, June 2021.

8. Bohan Li, **Yasheng Sun**, Jingxin Dong, Zheng Zhu, Jinming Liu, Xin Jin, Wenjun Zeng. One at a Time: Progressive Multi-Step Volumetric Probability Learning for Reliable 3D Scene Perception. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '24), pages 3028-3036, February 2024.

9. Bohan Li, **Yasheng Sun**, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, Wenjun Zeng. StereoScene: BEV-Assisted Stereo Matching Empowers 3D Semantic Scene Completion. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI '24), August 2024.