IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 March 2024, accepted 7 April 2024, date of publication 17 April 2024, date of current version 26 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3390182

RESEARCH ARTICLE

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

YASHENG SUN [1], WENQING CHU [2], HANG ZHOU[2], KAISIYUAN WANG[3], AND HIDEKI KOIKE [1]

[1]Tokyo Institute of Technology, Tokyo 152-8550, Japan
[2]Baidu Inc., Beijing 100085, China
[3]School of Electrical and Computer Engineering, The University of Sydney, Darlington, NSW 2008, Australia

Corresponding author: Yasheng Sun (sun.y.aj@m.titech.ac.jp)

ABSTRACT While considerable progress has been made in achieving accurate lip synchronization for 3D speech-driven talking face generation, the task of incorporating expressive facial detail synthesis aligned with the speaker's speaking status remains challenging. Existing efforts either focus on learning a dynamic talking head pose synchronized with speech rhythm or aim for stylized facial movements guided by external reference such as emotional labels or reference video clips. The former works often yield coarse alignment, neglecting the emotional nuances present in the audio content while the latter studies lead to unnatural applications, requiring manual style source selection by users. Our goal is to *directly leverage the inherent style information conveyed by human speech for generating an expressive talking face* that aligns with the speaking status. In this paper, we propose **AVI-Talking**, an **A**udio-**V**isual **I**nstruction system for expressive **Talking** face generation. This system harnesses the robust contextual reasoning and hallucination capability offered by Large Language Models (LLMs) to instruct the realistic synthesis of 3D talking faces. Instead of directly learning facial movements from human speech, our two-stage strategy involves the LLMs first comprehending audio information and generating instructions implying expressive facial details seamlessly corresponding to the speech. Subsequently, a diffusion-based generative network executes these instructions. This two-stage process, coupled with the incorporation of LLMs, enhances model interpretability and provides users with flexibility to comprehend instructions and specify desired operations or modifications. Specifically, given a speech clip, we first employ a Q-Former for contrastive alignment the speech features with visual instructions, which is then projected to input text embedding of LLMs. It functions as a prompting strategy, prompting LLMs to generate plausible visual instructions that encompass diverse facial details. In order to use these predicted instructions, a language-guided talking face generation system with disentangled latent space is delicately derived, where the speech content related lip movements and emotion correlated facial expressions are separately represented in *speech content space* and *content irrelevant space*. Additionally, we introduce a contrastive instruction-style alignment and diffusion technique within the content-irrelevant space to fully exploit the talking prior network for diverse instruction-following synthesis. Extensive experiments showcase the effectiveness of our approach in producing vivid talking faces with expressive facial movements and consistent emotional status.

INDEX TERMS Large language models, audio-visual instruction, diffusion model, expressive talking face generation, contrastive learning.

I. INTRODUCTION

Generating realistic 3D animations of human faces is crucial for a multitude of entertainment applications, encompassing digital human animation, visual dubbing in movies, and the creation of virtual avatars. To synthesize expressive speech-driven 3D talking face, previous work either 1) model the correlation between dynamic head poses and audio rhythm [1], [2] or 2) borrow an external representation [3], [4], [5] such as emotion labels or video clips as style reference

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

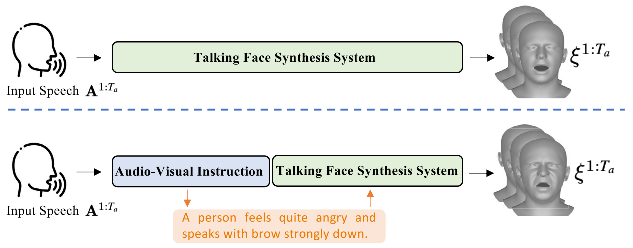

Input Speech $\mathbf{A}^{1:T_a}$
Talking Face Synthesis System
$\xi^{1:T_a}$
Input Speech $\mathbf{A}^{1:T_a}$
Audio-Visual Instruction
Talking Face Synthesis System
$\xi^{1:T_a}$
A person feels quite angry and speaks with brow strongly down.

FIGURE 1. In contrast to previous approaches that directly learn facial motions from speaker speech, our framework introduces an audio-visual instruction module achieved by LLMs to instruct the talking face synthesis network.

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 [1], [6], [7] 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 [3] 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.

To handle the above challenges, 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, 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.

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) [8], [9], we propose integrating LLMs as an agent [10] 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 [4]. 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.

Compared with previous studies, our main contributions in this work can be summarized as follows:

- 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.
- 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.

- 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.
- Experimental results validate the capability of AVI-Talking in generating vivid 3D talking faces with expressive facial details and a consistent emotional status.

II. RELATED WORK

Speech-driven facial animation holds diverse applications in the realms of computer vision and augmented/virtual reality. This has spurred a broad spectrum of research topics encompassing both 2D [11], [12], [13], [14], [15] and 3D [5], [16], [17], [18], [19], [20], [21], [22] facial synthesis. In the subsequent discussion, we delve into the most relevant studies in this field.

A. EXPRESSIVE 2D TALKING FACE GENERATION

Facial expressions play a pivotal role in the generation of natural talking heads [23]. Researchers [3], [24], [25], [26], [27], [28], [29], [30] attempt to synthesize vivid facial details while produce precise lip-synchronization.

Early approaches [31], [32], [33], [34], [35] represent expressions using a limited set of emotion labels encoded as one-hot representations. To capture nuanced talking face expressions, another slew of methodologies [3], [28], [29] incorporate reference videos as a more diverse stylistic source. While operating on RGB videos, these approaches rely on intricately designed disentanglement strategies. However, such practice often results in constrained expressiveness due to the inherent challenges of disentanglement. Meanwhile, these 2D animation stylized talking face methods have limited applicability in scenarios that demand 3D representations, such as in augmented reality (AR).

Instead of requiring users to seek out a stylized source, a more user-friendly approach involves directly exploiting speaking styles from the input audio. While some methods [26], [27], [36] derive 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 [1] or expressions [6] 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, allowing users to specify their desired styles [4], [37]. 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.

B. SPEECH-DRIVEN 3D TALKING HEAD GENERATION

Unlike 2D facial animation, which operates on RGB videos, 3D talking head generation employs speech-conditioned animation, utilizing geometric representations like the neural radius field (NerF) [38] or 3D parametric templates [19]. While methods such as [5], [19], [21], and [39] 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 [2] introduce a diffusion mechanism for its remarkable generative capability, yielding diverse high-quality synthesis results [40], [41]. 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.

C. LLM FOR CROSS-MODAL LEARNING

Large language models (LLMs) have demonstrated profound capabilities [42], [43] as remarkable reasoning engines in various language generation tasks, attributed to their emergent ability [44]. Diverse LLMs, such as OPT [45], LLaMA [46], and GLM [47], can be fine-tuned or instructed for various purposes [48]. Specifically, many studies attempt to construct LLMs proficient in multi-modal reasoning and actioning [8], leading to the emergence of MM-LLMs. Some studies point out that LLMs could even outperform diffusion models on standard image and video generation benchmarks [9]. In the pursuit of LLMs capable of handling both multi-modal input and output, some approaches explore employing LLMs as decision-makers [49] and utilizing existing off-the-shelf multi-modal encoders and decoders as tools for executing multi-modal input [50] and output [51], [52], [53].

Recent advancements in talking face generation have demonstrated the language model's capacity to generate multi-modal content [54] and synthesize facial motions [37], [55]. Typically, these approaches involve deriving special tokens for another modality and learning a projection layer to align them with language space [55]. 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

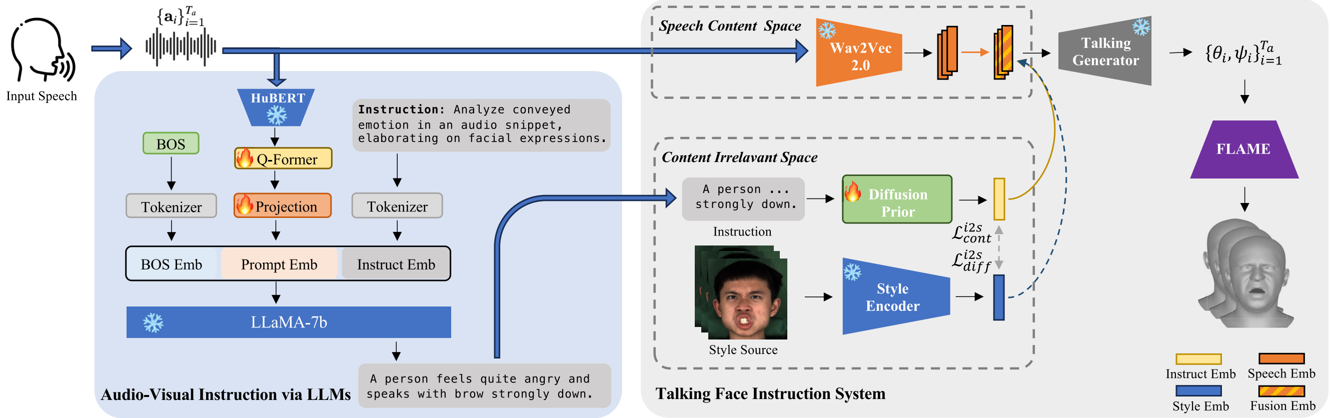

Input Speech
$\{\mathbf{a}_i\}_{i=1}^{T_a}$
HuBERT
Q-Former
Projection
BOS
Tokenizer
Tokenizer
Tokenizer
Instruction: Analyze conveyed emotion in an audio snippet, elaborating on facial expressions.
BOS Emb
Prompt Emb
Instruct Emb
LLaMA-7b
A person feels quite angry and speaks with brow strongly down.
Audio-Visual Instruction via LLMs
Speech Content Space
Wav2Vec 2.0
Talking Generator
$\{\theta_i, \psi_i\}_{i=1}^{T_a}$
FLAME
Content Irrelavant Space
A person ... strongly down.
Instruction
Diffusion Prior
$\mathcal{L}_{cont}^{i2s}$
$\mathcal{L}_{diff}^{i2s}$
Style Source
Style Encoder
Talking Face Instruction System
Instruct Emb
Speech Emb
Style Emb
Fusion Emb

FIGURE 2. The overall pipeline of our Audio-Visual Instruction Talking (AVI-Talking) Framework. 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 separately fed into the talking face instruction system to generate a time series of 3D parametric coefficients $\{\theta_i, \psi_i\}_{i=1}^{T_a}$.

eliminates the need for challenging cross-modal alignment procedures, which also provides enhanced explainability and flexibility to users.

III. METHODOLOGY

We present an **Audio-Visual Instruction** System for Expressive **Talking** Face Generation (**AVITalking**) that aims to achieve vivid facial expressions synthesis coherent with speaker speech status. The whole pipeline is depicted in Figure. 2. In this section, we start by briefly outlining the fundamentals of the parametric 3D face model and diffusion models (Sec. III-A). Subsequently, we present an overview of our pipeline (Sec. III-B). We then delve into the process of *audio-visual instruction* utilizing Large Language Models (LLMs) (Sec. III-C). Finally, we provide detailed description for *instruction-following talking face synthesis* (Sec. III-D).

A. PRELIMINARIES

1) PARAMETRIC 3D FACE MODEL

Animating a template mesh that encapsulates a 3D structural representation holds promise not only for AR/VR applications but also for facilitating 2D talking face synthesis [30]. However, the availability of 3D datasets capturing expressive facial movements is limited. Therefore, we employ a 3D reconstruction method [56] to convert video clips from 2D audio-visual datasets [4], [57] into 3D talking face datasets. Meanwhile, such practice provides both 2D images and 3D representations to enhance the training process.

Specifically, we adopt FLAME [58] as our template mesh. The FLAME model is a parametric 3D head model expressed as a function $M(\beta, \theta, \psi) \rightarrow (V, F)$, where the parameters consist of identity 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 $V \in \mathbb{R}^{n_v \times 3}$ and faces $F \in \mathbb{R}^{n_f \times 3}$, where n_v represents the number of vertices and n_f denotes the number of faces.

2) DIFFUSION MODEL

The goal of generative models is to learn a distribution that approximates real data distribution $q(x_0)$. The denoising diffusion probabilistic models (DDPMs) [59] present a multi-step progress to approximate $q(x_0)$ with $p_\theta(x_0)$ 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 $q(x_{1:T}|x_0)$. This process follows a Markov chain that progressively introduces Gaussian noise to the data samples. Formally,

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

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}). \tag{2}$$

Here, the constants β_t follow an increasing trend [59] 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,

$$p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \tag{3}$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)). \tag{4}$$

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 integrated into the network architecture. Formally,

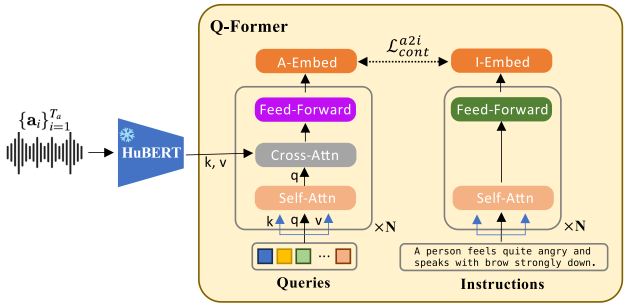

Q-Former
$\mathcal{L}^{a2i}_{cont}$
A-Embed
I-Embed
Feed-Forward
Feed-Forward
$\{\mathbf{a}_i\}^{T_a}_{i=1}$
HuBERT
k, v
Cross-Attn
q
Self-Attn
Self-Attn
k q v
×N
×N
Queries
A person feels quite angry and speaks with brow strongly down.
Instructions

FIGURE 3. The Q-Former architecture leverages the standard Perceiver network [63] to compress speech input to a fixed-length audio embedding $F^a_{si} \in \mathcal{R}^{q_a \times l}$. A contrastive loss $\mathcal{L}^{a2i}_{cont}$ is applied to encourage the queries extract audio representation that are most relevant to visual instructions.

the model parameters θ are optimized for all sampled timestamps t and $\boldsymbol{x}$ with the following objective:

$$\mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{x},t}[\|\boldsymbol{x} - \mathcal{G}_\theta(\boldsymbol{x}, t, \boldsymbol{c})\|^2]. \quad (5)$$

B. OVERALL PIPELINE

Given an audio clip, our objective is to animate a template mesh with synchronized lip movements and consistent facial expressions. Instead of directly learning to synthesize a talking face from speech, we propose integrating Large Language Models (LLMs) to instruct the synthesis process. As illustrated in Figure. 2, 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. Formally, our system accepts an input speech $A^{1:T_a} = \{a_i\}^{T_a}_{i=1}$ and aims to generate a time series of 3D parametric coefficients $\xi^{1:T_a} = \{\theta_i, \psi_i\}^{T_a}_{i=1}$.

C. AUDIO-VISUAL INSTRUCTION VIA LLMS

As illustrated on the left side of Figure 2, 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 [60] on Speech Emotion Recognition [61] (SER) tasks, we leverage HuBERT to encode the audio signal. Subsequently, a typical Q-Former [62], [63] architecture is employed to aggregate and extract speaking style information, bridging the gap between acoustic feature and visual facial descriptions. 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.

1) 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 [62], [63] to achieve this task.

As depicted in left side of Figure 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 $F^a_{si} \in \mathcal{R}^{q_a \times l}$ 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.

2) 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 [64] protocol to constrain the output of learned queries $F^a_{si} \in \mathcal{R}^{q_a \times l}$. 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 3. Its output embedding of $[CLS]$ token is $F^i_{si} \in \mathcal{R}^l$. Since there are q_a query embeddings, we average F^a_{si} across all queries to obtain the $\bar{F}^a_{si} \in \mathcal{R}^l$ and apply contrastive learning as follows:

$$\mathcal{L}^{a2i}_{cont} = -\log[\frac{\exp(\mathcal{D}(\bar{F}^a_{si}, F^i_{si}))}{\exp(\mathcal{D}(\bar{F}^a_{si}, F^i_{si})) + \sum_{j=1}^{N^-} \exp(\mathcal{D}(\bar{F}^a_{si}, F^{i-}_{si(j)}))}]. \quad (6)$$

The paired in-batch samples are regarded as positive samples $(\bar{F}^a_{si}, F^i_{si})$ while the unpaired N^- samples are taken as negative samples $(\bar{F}^a_{si}, F^{i-}_{si(j)})$. Here we opt for cosine distance $\mathcal{D}(F_1, F_2) = \frac{F_1^{\mathrm{T}} * F_2}{|F_1| \cdot |F_2|}$ as feature distance measurement.

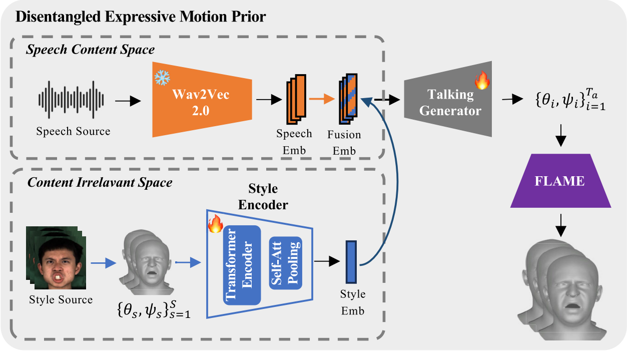


FIGURE 4. To establish a disentangled expressive motion prior, we learn two complementary latent spaces, *speech content* space and *content irrelevant* space. In *speech content* space, we represent lip movements related to speech content, while in the *content irrelevant* space, we capture facial expressions correlated with the speaking state.

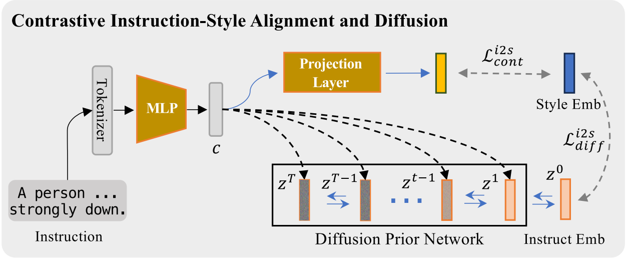

Contrastive Instruction-Style Alignment and Diffusion
Tokenizer
MLP
c
Projection Layer
$\mathcal{L}_{cont}^{i2s}$
Style Emb
$\mathcal{L}_{diff}^{i2s}$
A person ... strongly down.
Instruction
z^T
z^{T-1}
z^{t-1}
z^1
z^0
Diffusion Prior Network
Instruct Emb

FIGURE 5. Within the content irrelevant space, 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.

3) 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 2. Specifically, We follow the general text generation training paradigm [50] to learn this projection layer.

D. 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 illustrated on the right side of Figure 2. The movements of the lips and facial expressions exhibit a high degree of correlation with each other [3]. For example, specific pronunciations often convey relevant emotions. To address this correlation and potential entanglement, we propose initially training a disentangled talking prior [31], [65], wherein the speech content space and content irrelevant space are distinguished (shown in Figure 4). Subsequently, a diffusion prior module (shown in Figure 5) is devised to bridge the gap between instruction text and talking styles within the identified content irrelevant space.

1) DISENTANGLED EXPRESSIVE MOTION PRIOR

As depicted in Figure. 4, 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 [66] to encode the speaker audio $A^{1:T_a}$. 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 [3] 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 A^t in the *speech 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.

2) 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 [4]. As depicted in Figure 5, 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 [67]. Formally,

$$\mathcal{L}_{cont}^{i2s} = -\log[\frac{\exp(\mathcal{D}(F_{ci}^i, F_{ci}^s))}{\exp(\mathcal{D}(F_{ci}^i, F_{ci}^s)) + \sum_{j=1}^{N^-}\exp(\mathcal{D}(F_{ci}^i, F_{ci(j)}^{s-}))}]. \tag{7}$$

The 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 F_{ci}^s denotes its corresponding embedded style feature within the content-irrelevant space. The batch-wise (F_{ci}^i, F_{ci}^s) instruction and style feature pairs are taken as positive samples while the unpaired N^- instances $(F_{ci}^i, F_{ci(j)}^{s-})$ are considered as negatives samples. Similarly, we adopt cosine distance $\mathcal{D}(F_1, F_2) = \frac{F_1^{\mathbf{T}} * F_2}{|F_1| \cdot |F_2|}$ as feature distance measurement.

TABLE 1. The quantitative results on MeadText [4] and RAVEDESS [57]. For all approaches, we compare them under three metrics including FID [69], KID [70] and LSE-D [71]. Lower scores indicate better performance.

	MeadText [4]			RAVEDESS [57]		
Method	FID ↓	KID ↓	LSE-D ↓	FID ↓	KID ↓	LSE-D ↓
MeshTalk [65]	201.06	0.3601	10.51	134.47	0.2831	9.19
EmoTalk [5]	124.41	0.2118	**8.37**	122.95	0.1929	**8.51**
CodeTalker [21]	68.68	0.0658	8.38	46.90	0.0711	8.99
FaceFormer [19]	68.35	0.0611	9.08	47.78	0.0721	8.85
GT	-	-	9.36	-	-	9.05
AVI-Talking (w/o LLM)	12.91	0.0205	8.95	16.59	0.0259	8.56
AVI-Talking	**12.53**	**0.0190**	9.06	**15.94**	**0.0225**	8.81

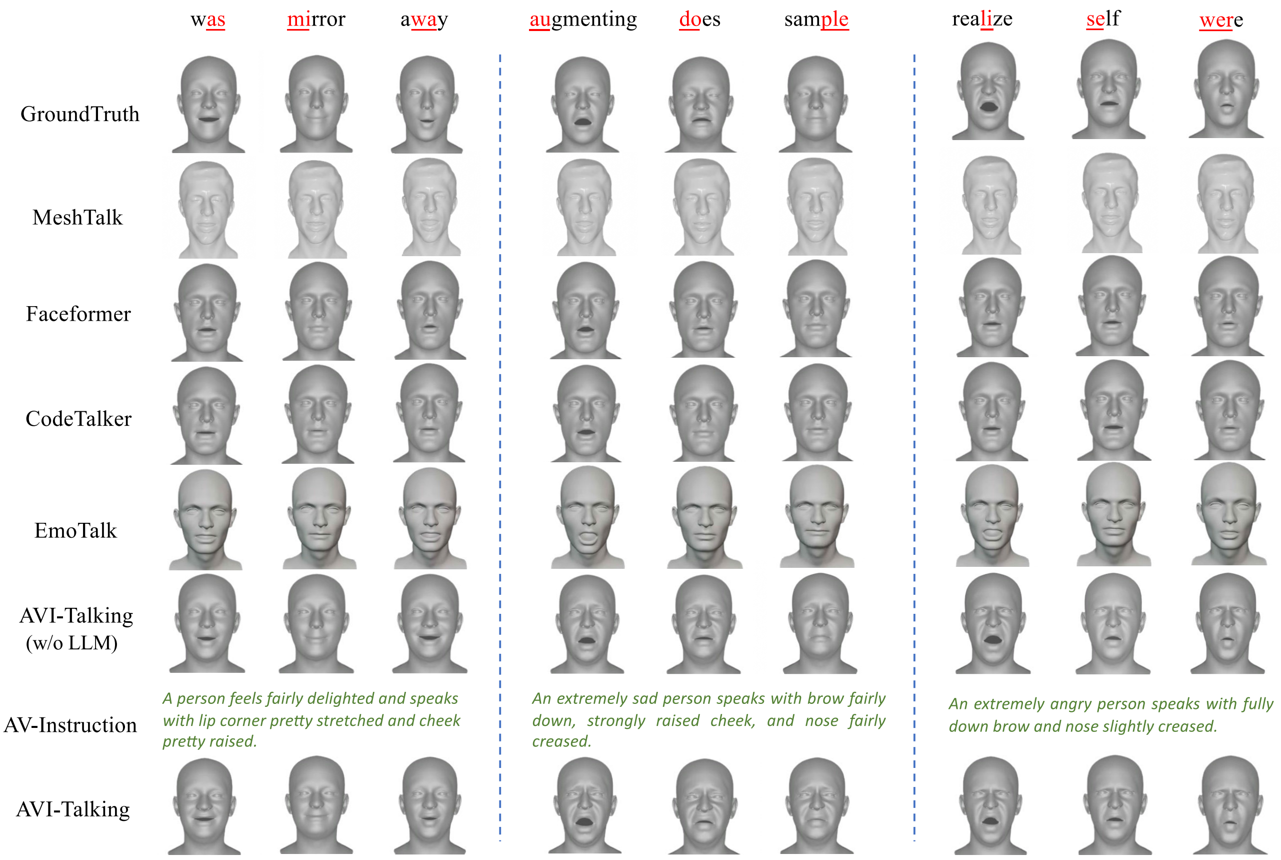

was
mirror
away
augmenting
does
sample
realize
self
were
GroundTruth
MeshTalk
Faceformer
CodeTalker
EmoTalk
AVI-Talking (w/o LLM)
AV-Instruction
A person feels fairly delighted and speaks with lip corner pretty stretched and cheek pretty raised.
An extremely sad person speaks with brow fairly down, strongly raised cheek, and nose fairly creased.
An extremely angry person speaks with fully down brow and nose slightly creased.
AVI-Talking

FIGURE 6. Qualitative Results. In the top row are ground truth videos. Our generated 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).

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 [68]. 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 z^t conditioned on the above representation $\mathbf{c}$. Instead of imposing error prediction formulation [59], we directly train the network to predict unnoised style embedding $\mathbf{z}$ from noised embedding z^t sampled at time step t. Formally,

$$\mathcal{L}^{i2s}_{diff} = \mathbb{E}_{z,t}[\|z - \mathcal{F}_\theta(z, t, \mathbf{c})\|^2] \tag{8}$$

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

$$\mathcal{L}^{i2s} = \mathcal{L}^{i2s}_{cont} + \lambda^{i2s}\mathcal{L}^{i2s}_{diff}, \tag{9}$$

TABLE 2. User study measured by Mean Opinion Scores. Larger is higher, with the maximum value to be 5.

MOS on \ Approach	MeshTalk [65]	EmoTalk [5]	CodeTalker [21]	**AVI-Talking (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**

where λ is the balancing coefficient. In our experiment, we empirically set it to 30, following a similar protocol to previous work [67].

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

We train both audio-visual instruction module and talking face instruction network on MeadText [4] dataset. Evaluation is conducted on test set of RAVEDESS [57] and MeadText. Since both datasets are made of RGB videos, we obtain reconstruction results by Emoca [56] and render the facial meshes as GT videos for comparison.

- **MeadText** [4]. This dataset is extended from Mead [35] dataset by labeling the speaker emotional status and facial action unit (FAU) with natural language descriptions. MEAD [35] is a high-quality emotional talking-face dataset, including recorded videos of different actors speaking with 8 different emotions at 3 intensity levels.
- **RAVEDESS** [57]. 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.

2) 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 [60] 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 [72], 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 [31] 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 [66] to extract the raw waveform and compress features with temporal convolutions following a similar protocol to EMOTE [31]. For speech style extraction, we follow the architecture design of StyleTalk [3] and leverage the linear styling network from EMOTE [31] 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 [67]'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 [73] 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.

TABLE 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 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**

3) COMPARISON METHODS

We compare our methods with state-of-the-art template-based models that support speech conditional 3D talking face generation, including MeshTalk [65], FaceFormer [19], CodeTalker [21], and EmoTalk [5].

MeshTalk [65] introduces a cross-modality disentanglement mechanism to generate realistic face animation. FaceFormer [19] devises a transformer-based architecture capable of synthesizing realistic 3D facial motions. CodeTalker [21] incorporates the codebook technique [74] to enhance the accuracy of lip movements. EmoTalk [5] 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).

B. QUANTITATIVE EVALUATION

1) 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*$_4$ [75], *METEOR* [76], *ROUGE*$_l$ [77], *CIDEr* [78] and *SPICE* [79].
- **3D Talking Face Synthesis.** To assess visual fidelity, we utilize standard GAN metrics: **FID** [69] and **KID** [70] on face regions of rendered images. Additionally, to evaluate generation diversity, we report **Diversity** scores [80], 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** [71] to evaluate lip synchronization performance.

2) EVALUATION RESULTS

Regarding the synthesis of talking faces, our study reports quantitative results for MeadText [4] and RAVEDESS [57] in Table 1. Notably, 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 [71], 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.

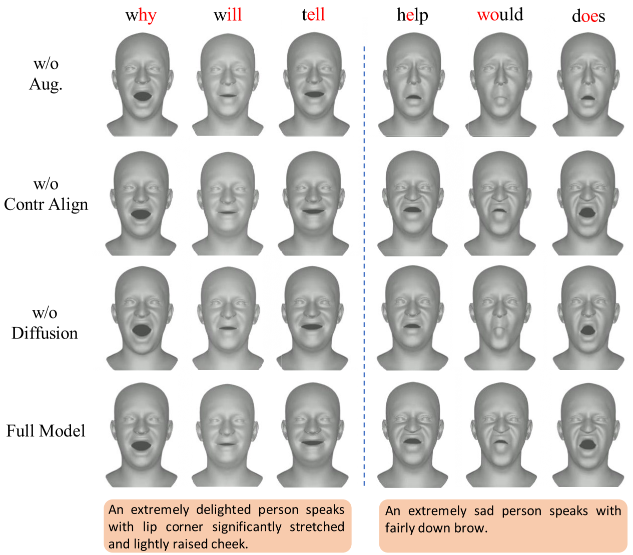

why
will
tell
help
would
does
w/o
Aug.
w/o
Contr Align
w/o
Diffusion
Full Model
An extremely delighted person speaks with lip corner significantly stretched and lightly raised cheek.
An extremely sad person speaks with fairly down brow.

FIGURE 7. Ablation Study. The bottom row illustrates the audio-visual instruction, while the rows above visualize the generation results across three key aspects of model design. Without diffusion, the model tends to produce conservative results, thereby inadequately raising the lip corner during smiling. Not utilizing data augmentation can result in sub-optimal convergence, failing to capture precise detailed facial movements.

C. QUALITATIVE EVALUATION

1) 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 6, we present comparison results of our method against previous state-of-the-art approaches in three cases. 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 [21] or Faceformer [19] may generate more natural pronunciation in expressionless states. However, when involving 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 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

2) 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.

The results are presented in Table 2. MeshTalk [65] scores the lowest across all aspects, possibly attributed to the architecture design of its naive UNet. By incorporating transformer blocks, EmoTalk [5] and CodeTalker [21] 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.

D. FURTHER ANALYSIS

1) 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.

a: 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 [62] and utilize image-grounded text generation loss for instruction generation. The numerical results on the MeadText dataset [4] are presented in Table 3. We find that without text data augmentation, the model tends to overfit to a sub-optimal point, leading to slightly worse performance. 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.

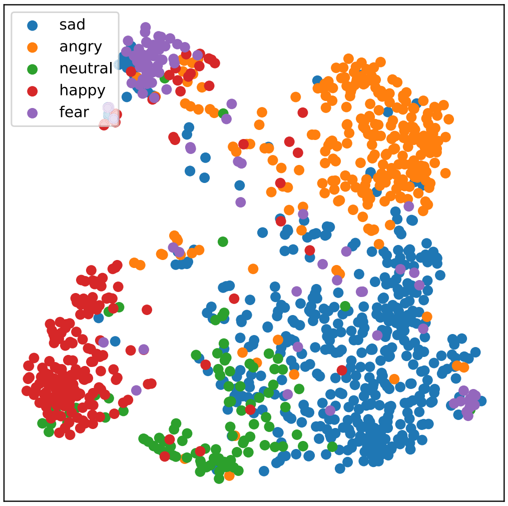
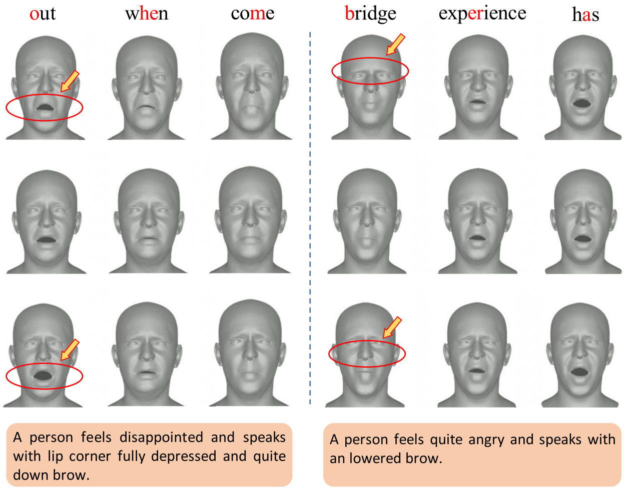


FIGURE 8. Visualizations of t-SNE embeddings derived from aligned speech features using Q-Former. The audio samples are from male utterances in the MeadText dataset, focusing on five typical speaking emotions. Notably, the aligned speech features corresponding to each specific emotion exhibit closely clustered patterns.



A person feels disappointed and speaks with lip corner fully depressed and quite down brow.

A person feels quite angry and speaks with an lowered brow.

FIGURE 9. Diverse generation results of the talking face instruction system are depicted. 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.

b: 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 [4] are demonstrated in Table 4, and visualization results are depicted in Figure 7. 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,

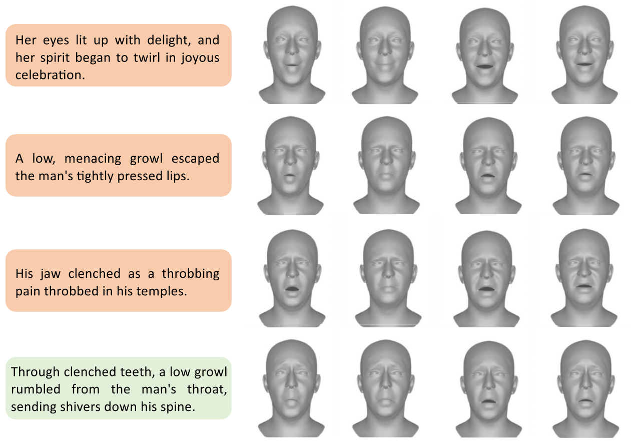

Her eyes lit up with delight, and her spirit began to twirl in joyous celebration.
A low, menacing growl escaped the man's tightly pressed lips.
His jaw clenched as a throbbing pain throbbed in his temples.
Through clenched teeth, a low growl rumbled from the man's throat, sending shivers down his spine.

FIGURE 10. **Visualization of Out-of-Distribution (OOD) results from the Talking Face Instruction System. 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.**

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 7 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.

2) 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. III-C, 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 8, 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.

3) GENERATION DIVERSITY OF TALKING FACE INSTRUCTION SYSTEM

In Table 4, we illustrate the pivotal role of diffusion strategy in enhancing generation diversity. Additionally, in Figure 9, 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.

4) OOD ANALYSIS OF TALKING FACE INSTRUCTION SYSTEM

To further assess the 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 10, 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.

V. CONCLUSION

In this paper, we propose **AVI-Talking**, an **Audio-Visual Instruction** system for expressive 3D **Talking** face generation. 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.

Limitations. Our model is currently trained on a labeled audio-visual instruction dataset. 1) We observe it exhibits insensitivity to certain specific speaking statuses. This phenomenon could be attributed to uneven data distribution, where certain speaking states are not adequately represented in the training dataset, making them challenging to discern from the speaker's speech. 2) The capabilities of the talking face synthesis network are limited to handling visual instructions closely aligned with the overall dataset distributions. As a consequence, for optimal instruction-following performance, users must provide instructions that closely resemble the predefined instructions.

Future Work. In this paper, we have investigated into specifying a pre-trained Large Language Model (LLM) for cross-modal audio-visual generation using finetuning techniques. Recent studies [81] 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 [9] 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.

Ethical Considerations. Our method has the potential to be exploited for malicious purposes, such as generating deepfakes, which can have detrimental effects on various aspects of society, including misinformation and privacy breaches. To mitigate this risk and ensure responsible usage, we have decided to limit the licensing of our model strictly to research purposes and will share it exclusively with the deepfake detection community. In addition to licensing restrictions, we will proactively incorporate robust watermarks into the generation process to facilitate the identification and tracking of deepfakes generated using our method. These watermarks will serve as an essential tool for forensic analysis, thereby enhancing the resilience of our technology against potential misuse.

REFERENCES

[1] L. Chen, G. Cui, C. Liu, Z. Li, Y. X. Z. Kou, and C. Xu, "Talking-head generation with rhythmic head motion," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 35–51.

[2] Z. Sun, T. Lv, S. Ye, M. G. Lin, J. Sheng, Y.-H. Wen, M. Yu, and Y.-J. Liu, "DiffPoseTalk: Speech-driven stylistic 3D facial animation and head pose generation via diffusion models," 2023, *arXiv:2310.00434*.

[3] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "StyleTalk: One-shot talking head generation with controllable speaking styles," 2023, *arXiv:2301.01081*.

[4] Y. Ma, S. Wang, Y. Ding, B. Ma, T. Lv, C. Fan, Z. Hu, Z. Deng, and X. Yu, "TalkCLIP: Talking head generation with text-guided expressive speaking styles," 2023, *arXiv:2304.00334*.

[5] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, and Z. Fan, "EmoTalk: Speech-driven emotional disentanglement for 3D face animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20687–20697.

[6] Z. Yu, Z. Yin, D. Zhou, D. Wang, F. Wong, and B. Wang, "Talking head generation with probabilistic audio-to-visual diffusion priors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7645–7655.

[7] Y. Ma, S. Zhang, J. Wang, X. Wang, Y. Zhang, and Z. Deng, "DreamTalk: When expressive talking head generation meets diffusion probabilistic models," 2023, *arXiv:2312.09767*.

[8] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "NExT-GPT: Any-to-any multimodal LLM," 2023, *arXiv:2309.05519*.

[9] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, B. Gong, M.-H. Yang, I. Essa, D. A. Ross, and L. Jiang, "Language model beats diffusion—Tokenizer is key to visual generation," in *Proc. The 12th Int. Conf. Learn. Represent.*, 2024, pp. 6–7. [Online]. Available: https://openreview.net/forum?id=gzqrANCF4g

[10] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, "A survey on large language model based autonomous agents," 2023, *arXiv:2308.11432*.

[11] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 716–731.

[12] O. Wiles, A. Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 670–686.

[13] Y. Sun, H. Zhou, Z. Liu, and H. Koike, "Speech2Talking-face: Inferring and driving a face with synchronized audio-visual representation," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, vol. 2, Aug. 2021, p. 4.

[14] Y. Sun, H. Zhou, K. Wang, Q. Wu, Z. Hong, J. Liu, E. Ding, J. Wang, Z. Liu, and K. Hideki, "Masked lip-sync prediction by audio-visual contextual exploitation in transformers," in *Proc. SIGGRAPH Asia Conf. Papers*, Nov. 2022, pp. 1–9.

[15] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4174–4184.

[16] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Aug. 2017.

[17] A. Richard, C. Lea, S. Ma, J. Gall, F. de la Torre, and Y. Sheikh, "Audio- and gaze-driven facial animation of codec avatars," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 41–50.

[18] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–2.

[19] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "FaceFormer: Speech-driven 3D facial animation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18749–18758.

[20] B. Thambiraja, I. Habibie, S. Aliakbarian, D. Cosker, C. Theobalt, and J. Thies, "Imitator: Personalized speech-driven 3D facial animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20621–20631.

[21] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "CodeTalker: Speech-driven 3D facial animation with discrete motion prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12780–12790.

[22] R. Huang, P. Lai, Y. Qin, and G. Li, "Parametric implicit face representation for audio-driven facial reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12759–12768.

[23] N. Sadoughi and C. Busso, "Speech-driven expressive talking lips with conditional sequential generative adversarial networks," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1031–1044, Oct. 2021.

[24] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1398–1413, May 2020.

[25] H. Wu, J. Jia, H. Wang, Y. Dou, C. Duan, and Q. Deng, "Imitating arbitrary talking style for realistic audio-driven talking face synthesis," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1478–1486.

[26] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14075–14084.

[27] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick, "Emotion-controllable generalized talking face generation," 2022, *arXiv:2205.01155*.

[28] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3377–3386.

[29] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao, "EAMM: One-shot emotional talking face via audio-based emotion-aware motion model," in *Special Interest Group Comput. Graph. Interact. Techn. Conf. Proc.*, Aug. 2022, pp. 1–10.

[30] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8652–8661.

[31] R. Daněček, K. Chhatre, S. Tripathi, Y. Wen, M. Black, and T. Bolkart, "Emotional speech-driven animation with content-emotion disentanglement," in *Proc. SIGGRAPH Asia Conf. Papers*, Dec. 2023, pp. 1–13.

[32] Y. Gan, Z. Yang, X. Yue, L. Sun, and Y. Yang, "Efficient emotional adaptation for audio-driven talking-head generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22634–22645.

[33] S. Gururani, A. Mallya, T.-C. Wang, R. Valle, and M.-Y. Liu, "SPACE: Speech-driven portrait animation with controllable expression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20914–20923.

[34] S. Tan, B. Ji, and Y. Pan, "EMMN: Emotional motion memory network for audio-driven emotional talking face generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22146–22156.

[35] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 700–717.

[36] C. Xu, J. Zhu, J. Zhang, Y. Han, W. Chu, Y. Tai, C. Wang, Z. Xie, and Y. Liu, "High-fidelity generalized emotional talking face generation with multi-modal emotion space learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6609–6619.

[37] D. Wang, B. Dai, Y. Deng, and B. Wang, "AgentAvatar: Disentangling planning, driving and rendering for photorealistic avatar agents," 2023, *arXiv:2311.17465*.

[38] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "AD-NeRF: Audio driven neural radiance fields for talking head synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5764–5774.

[39] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10093–10103.

[40] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," 2022, *arXiv:2209.14916*.

[41] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2022. [Online]. Available: https://openreview.net/forum?id=L7wzpQttNO

[42] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.

[43] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," 2022, *arXiv:2212.10403*.

[44] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," 2022, *arXiv:2206.07682*.

[45] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[46] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[47] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, Z. Liu, P. Zhang, Y. Dong, and J. Tang, "GLM-130b: An open bilingual pre-trained model," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–2. [Online]. Available: https://openreview.net/forum?id=-Aw0rrrPUF

[48] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. Red Hook, NY, USA: Curran Associates, 2022, pp. 27730–27744.

[49] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," 2023, *arXiv:2302.04761*.

[50] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.

[51] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu, Y. Ren, Z. Zhao, and S. Watanabe, "AudioGPT: Understanding and generating speech, music, sound, and talking head," 2023, *arXiv:2304.12995*.

[52] Y. Xu, H. Chen, J. Yu, Q. Huang, Z. Wu, S. Zhang, G. Li, Y. Luo, and R. Gu, "SECap: Speech emotion captioning with large language model," 2023, *arXiv:2312.10381*.

[53] Y. Sun, Y. Yang, H. Peng, Y. Shen, Y. Yang, H. Hu, L. Qiu, and H. Koike, "ImageBrush: Learning visual in-context instructions for exemplar-based image manipulation," 2023, *arXiv:2308.00906*.

[54] K. Zheng, X. He, and X. Eric Wang, "MiniGPT-5: Interleaved vision-and-language generation via generative vokens," 2023, *arXiv:2310.02239*.

[55] E. Ng, S. Subramanian, D. Klein, A. Kanazawa, T. Darrell, and S. Ginosar, "Can language models learn to listen?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10083–10093.

[56] R. Danecek, M. Black, and T. Bolkart, "EMOCA: Emotion driven monocular face capture and animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20311–20322.

[57] S. R. Livingstone and F. 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*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[58] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194:1–194:17, 2017, doi: 10.1145/3130800.3130813.

[59] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[60] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[61] O. Mohamed and S. A. Aly, "Arabic speech emotion recognition employing Wav2vec2.0 and Hubert based on BAVED dataset," 2021, *arXiv:2110.04425*.

[62] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[63] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23716–23736.

[64] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[65] A. Richard, M. Zollhöfer, Y. Wen, F. de la Torre, and Y. Sheikh, "MeshTalk: 3D face animation from speech using cross-modality disentanglement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1153–1162.

[66] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 12449–12460.

[67] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[68] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 17612–17625.

[69] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5–6.

[70] Z. Ren, Z. Pan, X. Zhou, and L. Kang, "Diffusion motion: Generate text-guided 3D human motion by diffusion model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[71] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 484–492.

[72] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, and J. E. Gonzalez. (Jun. 23, 2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90%* ChatGPT Quality*. Accessed: Apr. 14, 2023. [Online]. Available: https://vicuna.lmsys.org

[73] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 3–5.

[74] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," 2020, *arXiv:2012.09841*.

[75] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[76] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2007, pp. 65–72.

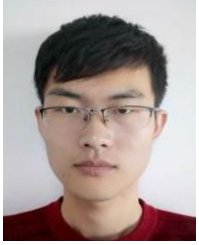
[77] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[78] Q. Wang and A. B. Chan, "Describing like humans: On diversity in image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4190–4198.

[79] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 382–398.

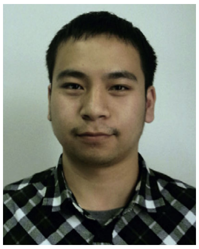
[80] S. Aneja, J. Thies, A. Dai, and M. Nießner, "FaceTalk: Audio-driven motion diffusion for neural parametric head models," 2023, *arXiv:2312.08459*.

[81] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2020, pp. 5–7.



YASHENG SUN received the B.E. degree from Nanjing University of Aeronautics and Astronautics and the M.E. degree from the School of Mechanical Engineering, Shanghai Jiao Tong University, China, in 2020. He is currently pursuing the Ph.D. degree in computer science with the School of Computing, Tokyo Institute of Technology, Japan.

His current research interests include cross-modal generation, 3D generative model, and stable diffusion model and its application in computer vision.



WENQING CHU received the B.E. degree in computer science and technology from Huazhong University of Science and Technology, in 2014, and the Ph.D. degree in computer science from Zhejiang University, in 2019.

He is currently a Researcher with Baidu Inc. His research interests include machine learning, computer vision, and data mining.

HANG ZHOU received the bachelor's degree in acoustics from the School of Physics, Nanjing University (NJU), in 2017, and the Ph.D. degree from the Multimedia Laboratory (MMLab), The Chinese University of Hong Kong, in 2021, supervised by Prof. Xiaogang Wang.

He is currently a Researcher with Baidu Inc. His research interests include deep learning and its applications on audio-visual learning, image/video generation, and virtual human creations.

KAISIYUAN WANG received the B.E. and M.E. degrees from the School of Electrical Engineering, Harbin Institute of Technology, China, in 2016 and 2018, respectively, and the Ph.D. degree from the School of Electrical and Information Engineering, The University of Sydney, in 2023.

While pursuing the Ph.D. degree, he was an Intern with Baidu Inc., China. His current research interests include audio/video-driven portrait synthesis and low-level tasks in 3D computer vision.

HIDEKI KOIKE received the B.E., M.E., and Dr. (Eng.) degrees from The University of Tokyo, in 1986, 1988, and 1991, respectively. He was with the University of Electro-Communications, Tokyo. In 2014, he joined Tokyo Institute of Technology, Japan, where he is currently a Professor with the School of Computing. His research interests include vision-based human–computer interaction, human augmentation, information visualization, and usable security.