InstructAvatar: Text-Guided Emotion and Motion Control for Avatar Generation

**Yuchi Wang, Junliang Guo*, Jianhong Bai, Runyi Yu, Tianyu He,
Xu Tan, Xu Sun*, Jiang Bian**

National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
{wangyuchi, ingrid_yu}@stu.pku.edu.cn, xusun@pku.edu.cn

Abstract

Recent talking avatar generation models have made strides in achieving realistic and accurate lip synchronization with the audio, but often fall short in controlling and conveying detailed expressions and emotions of the avatar, making the generated video less vivid and controllable. In this paper, we propose a text-guided approach for generating emotionally expressive 2D avatars, offering fine-grained control, improved interactivity and generalizability to the resulting video. Our framework, named InstructAvatar, leverages a natural language interface to control the emotion as well as the facial motion of avatars. Technically, we utilize GPT-4V to design an automatic annotation pipeline, constructing an instruction-video paired training dataset. This is combined with a novel two-branch diffusion-based generator to predict avatars using both audio and text instructions simultaneously. Experimental results demonstrate that InstructAvatar produces results that align well with both conditions, and outperforms existing methods in fine-grained emotion control, lip-sync quality, and naturalness.

Demo — https://wangyuchi369.github.io/InstructAvatar/
Extended version — https://arxiv.org/abs/2405.15758

1 Introduction

Avatar generation has recently gained significant attention due to its broad applicability in film production, gaming, video conferencing, and various other domains. The primary objective of this technology is to animate portraits with synchronized speech audio. While previous studies have achieved impressive lip synchronization and head pose prediction (Tian et al. 2024; Sun et al. 2023; He et al. 2024; Zhang et al. 2023b), effectively conveying and controlling detailed expressions and motions remains a challenge, resulting in less vivid and authentic videos. Some previous studies have attempted to integrate emotional information through either labels (Wang et al. 2020; Zhai et al. 2023; Tan, Ji, and Pan 2023) or example videos (Ma et al. 2023c,b; Zhang et al. 2023a). However, they continue to face challenges related to limited flexibility and controllability.

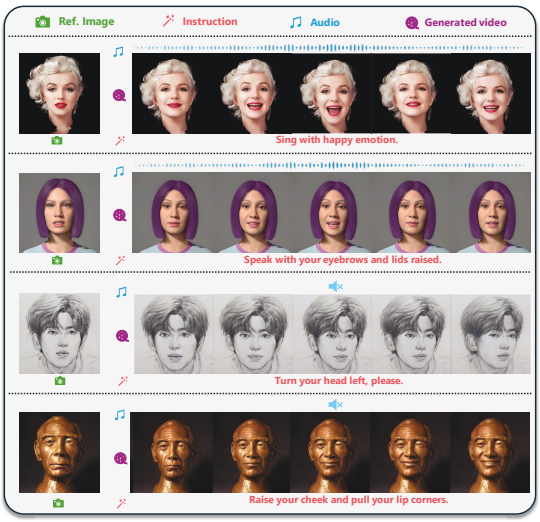

Figure 1: InstructAvatar enables emotional talking face generation through a flexible natural language interface (top 2 rows). Moreover, it supports direct control of facial motion and expression without relying on audio cues, a feature absent in previous studies (bottom 2 rows).

As a seamless interface bridging the gap between humans and computers, the textual prompt stands out as a natural solution to tackle the limitations of previous approaches, providing a versatile array of controls that encompass desired motions and expressions for the avatar. Therefore, we steer towards a textual instruction-based talking avatar generation model named InstructAvatar, to directly emit talking face video, which offers three key advantages: **(1)** Enhanced control over fine-grained details rather than just the overall style; **(2)** Improved generalizability compared to limited emotion or style categories; and **(3)** Enhanced interactivity and user-friendliness. As illustrated in Fig. 1, our framework enables **text-guided emotional talking face generation** with fine-grained control and allows for **facial motion and expression control** without relying on audio cues.

To accomplish this, we meticulously design our algorithm, taking into account both data and model architecture considerations. For data preparation, existing datasets (Wang et al. 2020; Cao et al. 2014) only offer tag-

*Junliang Guo and Xu Sun are corresponding authors.

level emotion annotations. To capture fine-grained facial details, we utilize Action Units (AUs) (Ekman and Friesen 1978) to describe facial muscle movements. To expand the range of emotion types and facial details, we employ a multi-modal language model (MLLM) like GPT-4V (OpenAI 2023) to paraphrase emotion types and AUs into natural textual descriptions. This process generates a dataset containing detailed emotion and motion descriptions. Regarding the model architecture of InstructAvatar, we propose a diffusion model with two branches of cross-attentions to incorporate different types of instructions while generating the talking video, i.e., the emotion instructions that are high-level throughout the entire video, and the facial motions that are dynamic over timestamps. Additionally, novel techniques such as zero-convolution gate are proposed to stabilize the training and enhance the guidance.

For experiments, we propose several tailored evaluation metrics to justify the model's performance on fine-grained facial emotion and motion control. Experimental results demonstrate that: (1) InstructAvatar exhibits significant improvements in emotion control, lip-sync quality, and naturalness compared to previous baselines. (2) Notably, by leveraging the MLLM, our model features a natural language interface, enabling it to receive a much wider range of instructions. (3) Additional experimental results indicate that our model can, for the first time, effectively animate avatars directly without audio. In summary, the contributions of our paper are as follows:

(1) Leveraging the MLLM, we introduce InstructAvatar, a diffusion-based avatar generation model featuring a fine-grained natural language interface. It showcases superior flexibility, control effectiveness, and naturalness compared to previous methods. To our best knowledge, it is the first text-guided 2D-based talking face generation framework.

(2) Through meticulous design, like the implementation of a two-branch cross-attention mechanism, we integrate text-guided facial motion control into our unified framework, further enhancing the scope of avatar control.

(3) We annotate an instruction-video dataset and establish an evaluation pipeline for the fine-grained emotional talking video generation task, which may facilitate further research.

2 Related Works

Talking head generation aims to generate a video in which an avatar speaks the provided audio. This task can be broadly categorized into video-driven and audio-driven approaches. In video-driven methods (Wang, Mallya, and Liu 2021a; Zhang et al. 2022; Tripathy, Kannala, and Rahtu 2021), the movement of a portrait is generated based on another driving video, while in audio-driven methods (Zhang et al. 2023b; Tian et al. 2024; He et al. 2024; Wang et al. 2024; Xu et al. 2024), motion is predicted directly from audio inputs. To tackle the challenge of learning facial motion representations, previous talking head models have often relied on domain priors like warping-based transformations (Wang et al. 2021; Zhou et al. 2020; Liu et al. 2022) or 3D Morphable Models (Zhang et al. 2021, 2023b). Recently, He et al. (2024) proposed a disentangled motion and appearance architecture and collected a large-scale dataset to directly learn the data distribution, thereby further enhancing the naturalness and diversity of the generated avatars.

Acknowledging the constraints of prior efforts that often yield emotionless avatars, there has been growing interest in injecting emotions into talking face generation. For instance, MEAD (Wang et al. 2020) represents emotion using a one-hot vector, while EAT (Gan et al. 2023) employs a mapping network to extract emotion guidance through a latent code. EAMM (Ji et al. 2022) represents the facial dynamics of reference emotional video as displacements to motion representations. PD-FGC (Wang et al. 2023b) disentangles control over specific facial organs with emotional expression, and StyleTalk (Ma et al. 2023b) develops a style encoder to extract the style of a reference video. These methods either support a limited range of coarse emotion types (Gan et al. 2023; Tan, Ji, and Pan 2023; Feng et al. 2024; Liang and Lu 2024; Liu et al. 2024) or necessitate users to seek out another desired style video (Wang et al. 2023b; Ji et al. 2022; Tan et al. 2024; Zhang et al. 2023a), limiting the flexibility and controllability of the generated avatars.

More recently, some endeavors have aimed to incorporate text as an emotion control signal (Wang et al. 2023a; Zhao et al. 2024; Zhong et al. 2023; Ma et al. 2023a; Tan, Ji, and Pan 2024; Sun et al. 2024). However, they typically utilize text to generate emotional talking 3D animations, which requires an external renderer to convert these animations into real talking videos. Although this animation-based approach may alleviate difficulty by introducing more domain priors, it inherently leads to indirect controls and restricted diversity (He et al. 2024). In this paper, we propose to directly learn the distribution of talking videos and enable fine-grained controlling with textual prompts, improving the naturalness and controllability of generation results.

3 Methodology

3.1 Overview

Given a sequence of audio clips $\boldsymbol{A} = [a_1, a_2, ..., a_N]$, one portrait image $\boldsymbol{I}$, and the text instruction $\boldsymbol{T}$, our model is tasked with animating the portrait to utter the audio with the target style represented by the instruction. In other words, we aim to learn a mapping to generate a video $\boldsymbol{V} = \mathcal{F}(\boldsymbol{A}, \boldsymbol{I}, \boldsymbol{T})$. As illustrated in Fig. 2, we decompose $\mathcal{F}$ into two parts: variational autoencoder (VAE) $\mathcal{H}$ and diffusion-based motion generator $\mathcal{G}$. The VAE follows the approach outlined in (He et al. 2024) to disentangle motion information from appearance, which means that we can derive $\boldsymbol{V} = \mathcal{H}(\boldsymbol{M}, \boldsymbol{I})$, where $\boldsymbol{M}$ represents purely motion information and $\boldsymbol{I}$ is the provided portrait. More details about VAE and motion latent $\boldsymbol{M}$ can be found in the Appendix. Now, we can focus on learning the motion generator conditioned on audio and textual instructions, i.e., $\boldsymbol{M} = \mathcal{G}(\boldsymbol{A}, \boldsymbol{T})$. In the following sections, we will detail how to obtain fluent, diverse, and fine-grained text instructions $\boldsymbol{T}$ by MLLM, as well as how to design a diffusion model-based text-guided motion generator $\mathcal{G}$.

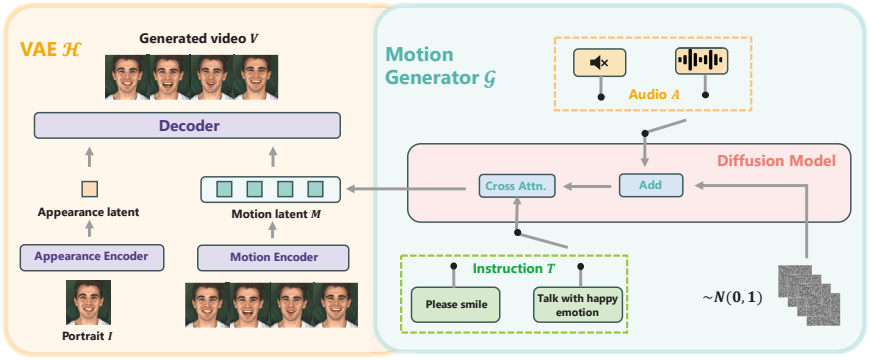

Figure 2: Method Overview: The InstructAvatar consists of two components: VAE $\mathcal{H}$ to disentangle motion information from the video and a motion generator $\mathcal{G}$ to generate the motion latent conditioned on audio and instruction.

3.2 Construct Natural and Diverse Text Instructions by MLLM

To utilize natural language as the interactive interface, it is essential to construct a dataset containing text-expression pairs. However, existing emotional talking datasets (Wang et al. 2020; Cao et al. 2014) typically provide only tag-level annotations for talking videos, offering limited emotion categories such as "happy" or "sad" along with their corresponding videos. To diversify and refine these labels into text instructions, we use action units (AUs) to obtain fine-grained facial information and employ MLLM like GPT-4V to generate open-vocabulary natural text instructions $\boldsymbol{T}$, as illustrated in Fig. 3.

Emotion Label Extension Firstly, we adopt a convenient method to convert emotion labels into sentences by utilizing predefined templates. Specifically, we prompt ChatGPT to generate 60 templates, such as "Talk with `[EMO]` emotion". We then randomly select a synonym from a predefined table and then substitute the placeholder `[EMO]` in the template, producing expressions like "Talk with delighted emotion". Additionally, we also utilize the emotion intensity information of the dataset. For instance, for videos with high emotion intensity, we may add adverbs like "extremely" to modify emotions, resulting in "extremely delighted".

Action Unit Extraction The previous method for emotion label-based extension tends to provide coarse and high-level annotations. To enable fine-grained control of facial expressions, we turn to the Action Units (AUs) (Ekman and Friesen 1978) to describe facial muscle movements, allowing for a detailed description of the local states of a talking face, as shown in Fig. 3. However, AU detection is typically performed on images. Therefore, we randomly select three frames from a video and employ an off-the-shelf AU detection model (Luo et al. 2022) to extract AUs from these images. We then take the intersection of predicted action units, considering that significant facial states corresponding to a specific emotion are likely to be consistent throughout the entire talking video.

MLLM Paraphrase The core procedure of data construction involves utilizing MLLM. Currently, the obtained information is relatively raw. For example, the action units obtained in the previous step are represented in a relatively incomprehensible form, such as "lid_tightner". More importantly, both the action units and emotion types are closed-vocabulary. To transform them into more human-friendly and open-vocabulary text instructions, we leverage large language models' powerful paraphrasing capabilities. We prompt GPT-4V (OpenAI 2023) to combine these action units into coherent sentences. Additionally, utilizing its vision capabilities, we provide GPT-4V with a frame extracted from the video and allow it to edit the action units if it disagrees with those extracted by the off-the-shelf model. We then ask GPT-4V to generate a few diverse sentences and randomly select one during training. Then we combine these sentences describing fine-grained details with the previously obtained sentence containing overall emotion. For training, we randomly select parts or use the whole sentences and further paraphrase some of them with ChatGPT.

3.3 Text-Guided Motion Generator

We leverage the diffusion model (Ho, Jain, and Abbeel 2020) as our text-guided motion generator to learn $\boldsymbol{M} = \mathcal{G}(\boldsymbol{A}, \boldsymbol{T})$ mentioned in Sec. 3.1. We use Conformer (Gulati et al. 2020) as our diffusion model backbone. The details of our motion generator is illustrated in Fig. 4.

Basics for Diffusion Models The diffusion model is designed to fit a distribution. Basically, it is divided into two phases: the forward diffusion process and the backward denoising process. Given a data point sampled from a distribution $x_0 \sim q(x)$, we define a forward process in which Gaussian noise is incrementally added to the sample, generating a sequence of noisy samples $x_1, ..., x_T$. The noise scales are controlled by $\beta_t \in (0, 1)$, and the density is expressed as $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$. Based on (Ho, Jain, and Abbeel 2020), we can sample at any arbitrary time step in a closed form: $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, \sqrt{1-\bar{\alpha}_t}\mathbf{I})$, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. For the reverse process, a simple learning objective is $L_{\text{simple}} = \sum_{t=1}^{T} \mathbb{E}_q \left[||\epsilon_t(x_t, x_0) - \epsilon_\theta(x_t, t)||^2 \right]$, where ϵ_t is the noise added in original data x_0 and ϵ_θ is learnable network. Recently, researchers tend to use an even sim-

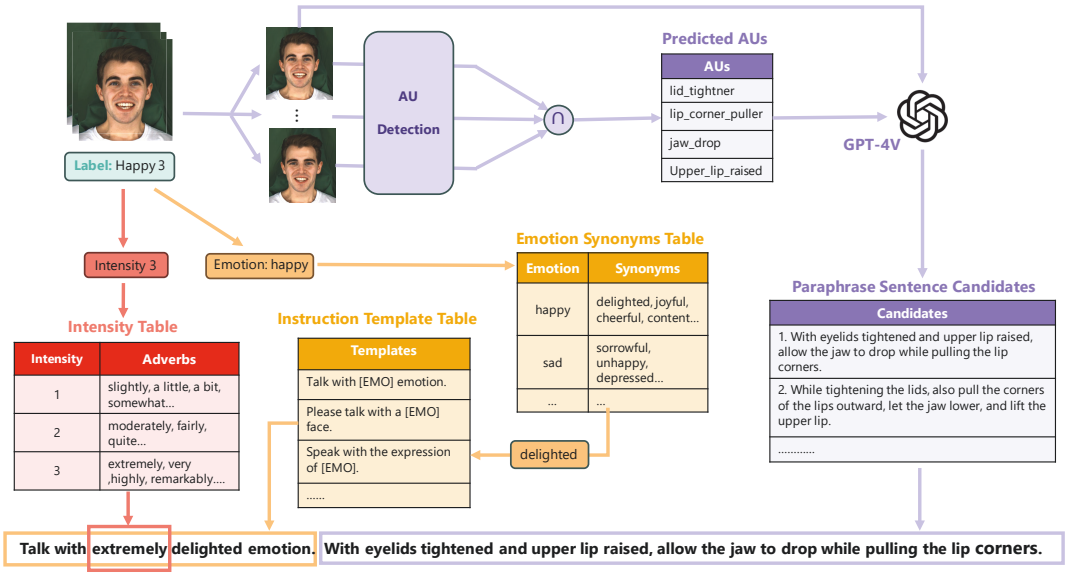

Figure 3: We extend the emotion label using a predefined template and incorporate intensity information by modifying the emotion with an adverb representing the degree. For fine-grained control, we extract the AUs and then prompt GPT-4V to paraphrase them into a sentence.

pler strategy to train a network to predict x_0 directly, with the loss function defined as $L = ||x_0 - f_\theta(x_t, t)||$, which is also applied in our model. During inference, following DDIM (Song, Meng, and Ermon 2020), we start from a Gaussian noise and iteratively denoise it to get predicted $\hat{x}_0$.

Audio-aware Input Block Basically, following the classical strategy of diffusion models mentioned above, we will train the denoising block to recover noised motion latent $\boldsymbol{M}_t$ to predicted $\hat{\boldsymbol{M}}_0$, denoted as $\hat{\boldsymbol{M}}_0 = f(\boldsymbol{M}_t, t, \boldsymbol{A}, \boldsymbol{T})$.

To incorporate audio information into the denoising process, we first normalize the speech to an appropriate amplitude range and then apply a denoiser (Defossez, Synnaeve, and Adi 2020) to reduce background noise. Subsequently, we utilize Wave2Vec 2.0 (Baevski et al. 2020) as audio encoder $\mathcal{W}$ to extract audio features. As a special case, for facial motion control absent of audio, we use pseudo empty audio with zero amplitude and a length aligned with the ground truth video. Given that audio and motion sequences are aligned in the temporal semantic, we opt to element-wise add the audio features $\boldsymbol{A}$ to the motion latent vector $\boldsymbol{M}$. In summary, we obtain the audio-aware noisy latent $\boldsymbol{M}_t^{\boldsymbol{A}}$ by:

$$\boldsymbol{M}_t^{\boldsymbol{A}} = \begin{cases} \boldsymbol{M}_t \oplus \mathcal{W}(\boldsymbol{A}) & \text{Emotional talking control.} \\ \boldsymbol{M}_t \oplus \mathcal{W}(\varnothing) & \text{Facial motion control.} \end{cases}$$

Two-branch Text-aware Denoising Block Now we would use cascaded denoising blocks to denoise $\boldsymbol{M}_t^{\boldsymbol{A}}$ to the data $\hat{\boldsymbol{M}}_0$. A key component of our architecture is injecting text instruction information into the denoising procedure. To encode the instructions $\boldsymbol{T}$ obtained in Sec. 3.2, we leverage the CLIP (Radford et al. 2021) text encoder $\mathcal{C}$, which has been proven for its powerful cross-modality alignment ability and semantic generalization ability. We employ a cross-attention mechanism to incorporate text information, where the hidden states in the Conformer layer act as queries, and the text representation serves as keys and values.

There exist some differences between emotion and motion controls. For the emotion, the text provides style guidance throughout the entire video, ensuring that the avatar maintains the desired emotion consistently. However, this is not the case for facial motion instructions, which usually describe gradually achieved actions and transitions over time. For example, a person may gradually turn his head when receiving corresponding motion instructions. In response, we split the data flow into two branches in each denoising block when incorporating text information. For the emotion branch, we use the `[EOS]` token from the CLIP text encoder, which encapsulates overall information about the instruction. For the motion branch, we utilize the hidden states of all tokens from the last layer of the CLIP text encoder to capture more detailed and dynamic information. Additionally, we introduce different Adapters (Gao et al. 2024) $\mathcal{A}_e, \mathcal{A}_m$ to better align the distributions of these two spaces with the space that facilitates the diffusion model's learning process. Overall, the instruction representation $\mathrm{Rep}(\boldsymbol{T})$ to inject into the denoising procedure for text-guided emotional talking and text-guided facial motion control could be summarized as:

$$\mathrm{Rep}(\boldsymbol{T}) = \begin{cases} \mathcal{A}_e(\mathcal{C}(\boldsymbol{T})_{[EOS]}) & \text{Emotional talking control.} \\ \mathcal{A}_m(\mathcal{C}(\boldsymbol{T})_{all}) & \text{Facial motion control.} \end{cases}$$

In addition, to provide the generator with facial shape information, following (He et al. 2024), we randomly select a frame from the motion latent $\boldsymbol{M}_0$ as the key frame latent $\boldsymbol{M}_k$ and inject it into the denoising process through cross-attention. It's noteworthy that for emotional talking, this se-

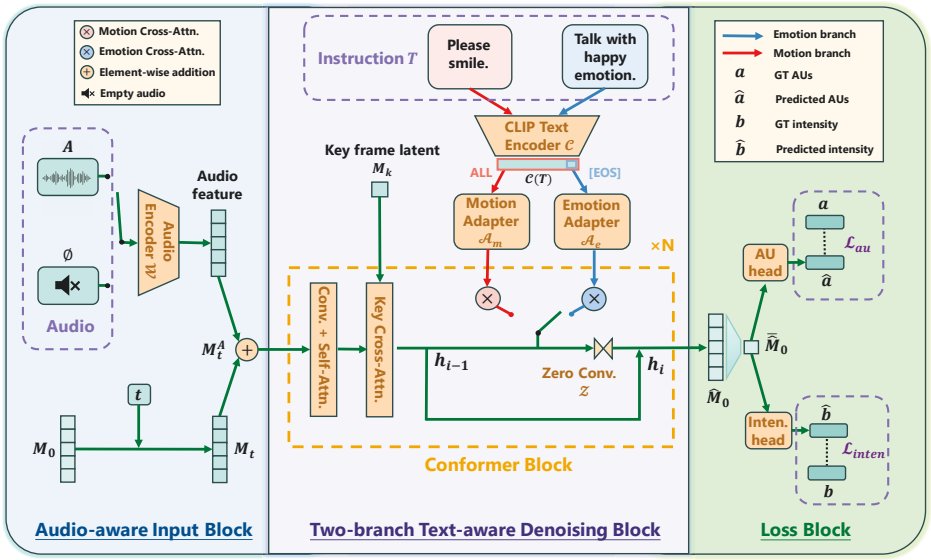

Figure 4: Details of motion generator. To denoise noisy motion latent, firstly we element-wise add the audio feature into it. Then, in each denoising block, we design a two-branch cross-attention module to inject emotion and motion control into the model. Lastly, we also incorporate AU and intensity losses to encourage the model to learn them.

lected key frame may inadvertently leak emotional information. Therefore, we substitute this frame with a frame from another emotional type video featuring the same person.

Zero Convolution Mechanism for Text Condition To leverage the abundant knowledge obtained from previous talking head models, we initialize our model from pretrained emotion-unaware models. Therefore, directly inserting text instructions would significantly reduce the expressiveness of the previously learned parameters. An insight is that an emotional talking video could transform from a neutral expression video gradually. Therefore, inspired by (Zhang, Rao, and Agrawala 2023), we tailor a zero convolution mechanism for text conditioning. Specifically, suppose we have hidden states h_{i-1} before entering the cross-attention module with text instructions. We then use $h_i = h_{i-1} + \mathcal{Z}(\text{Cross-Attn}(h, \text{Rep}(\boldsymbol{T})))$ to get the next hidden states. In this formula, $\mathcal{Z}$ is a zero convolution operation where a 1-dimensional convolutional kernel moves along the hidden states dimension, with both the weight and bias initialized to zero. Therefore, at the start of training, $h_i = h_{i-1}$, which corresponds to the no instruction setting. Consequently, the zero convolution layer serves as a gate to slowly inject text control instructions into the pre-trained talking face model, stabilizing the training process and leveraging the abundant knowledge obtained in previous emotion-unaware models.

3.4 Training and Inference Pipelines

Loss Definition After cascaded denoising blocks, we obtain the predicted motion latent $\hat{\boldsymbol{M}}_0$. The most intuitive loss is the distance between the predicted motion latent and the ground truth $\boldsymbol{M}_0$, expressed as $\mathcal{L}_{mse} = \|\hat{\boldsymbol{M}}_0 - \boldsymbol{M}_0\|_2^2$. Additionally, to enforce the model to pay attention to action units and emotion intensity, we jointly train two classifier heads. We perform mean pooling in the temporal dimension followed by two-layer MLPs to extract information from $\hat{\boldsymbol{M}}_0$. $\mathcal{L}_{au}$ is calculated using the binary cross-entropy (BCE) loss, treating it as a multi-label classification problem over predicted AU logits $\hat{\boldsymbol{a}} \in \mathbb{R}^M$ and ground-truth labels $\boldsymbol{a}$, given by: $\mathcal{L}_{au} = -\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{a} \odot \log(\hat{\boldsymbol{a}}) + (1-\boldsymbol{a}) \odot \log(1-\hat{\boldsymbol{a}})\|_1$, where M is the number of action units and N is the sample number. For the emotion intensity loss, which is a three-classification problem, we use the standard cross-entropy loss $\mathcal{L}_{inten} = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{b}\log(\hat{\boldsymbol{b}})$, where $\hat{\boldsymbol{b}} \in \mathbb{R}^3$ represents predicted logits and $\boldsymbol{b}$ represents the corresponding label. Moreover, following the approach in (He et al. 2024), we additionally train a head pose predictor and use another mean squared error (MSE) loss $\mathcal{L}_{pose} = \|\hat{\boldsymbol{P}} - \boldsymbol{P}\|_2^2$ to measure the predicted pose and the ground truth. In summary, our loss is defined as :

$$L = \mathcal{L}_{mse} + \lambda_{pose}\mathcal{L}_{pose} + \lambda_{au}\mathcal{L}_{au} + \lambda_{inten}\mathcal{L}_{inten}$$

where λ_{pose}, λ_{au}, λ_{inten} are hyperparameters.

Inference Pipeline During inference, we begin by sampling a Gaussian noise $\boldsymbol{M}_T$ to initialize the motion latent. The audio and instructions are provided by the user. We employ the VAE motion encoder (He et al. 2024) to encode the user-provided portrait $\boldsymbol{I}$, resulting in the keyframe motion latent $\boldsymbol{M}_k$. Subsequently, we iteratively denoise $\boldsymbol{M}_T$ using our trained denoising network, following the DDIM (Song, Meng, and Ermon 2020). Finally, we obtain the predicted motion latent $\hat{\boldsymbol{M}}$, and utilizing the VAE decoder $\mathcal{H}$, we generate the RGB video $\hat{\boldsymbol{V}} = \mathcal{H}(\hat{\boldsymbol{M}}, \boldsymbol{I})$.

4 Experiments

We evaluate our model for both emotional talking control and facial motion control. For emotional talking control, the

Method	$AU_{F1}\uparrow$	$AU_{Emo}\uparrow$	FID↓	$Sync_D\downarrow$	Emo.↑	Lip.↑	Jit.↑	Nat.↑	Guid.
GAIA (He et al. 2024)	0.549/0.185	0.352/0.048	52.716/–	9.542/9.776	3.02/1.83	4.63/4.51	4.57/4.55	4.31/4.47	-
MakeItTalk (Zhou et al. 2020)	0.588/0.220	0.405/0.065	47.269/–	11.291/10.059	3.88/2.36	3.49/3.62	3.99/3.68	3.82/3.46	-
EAT (Gan et al. 2023)	0.648/0.542	0.495/0.319	57.379/–	**8.757/8.962**	4.48/4.18	4.27/4.18	4.40/3.27	4.14/4.20	Label
StyleTalk (Ma et al. 2023b)	0.694/0.499	**0.593**/0.278	75.783/–	12.287/12.388	4.53/4.12	4.17/2.26	3.43/1.88	3.18/2.04	Video
DreamTalk (Ma et al. 2023c)	0.711/0.513	0.548/0.301	85.291/–	11.967/10.967	4.59/4.07	3.74/2.39	3.95/2.21	3.72/2.26	Video
InstructAvatar (Ours)	**0.738/0.552**	0.566/**0.324**	**44.593**/–	9.412/9.653	**4.64/4.52**	**4.74/4.59**	**4.88/4.68**	**4.63/4.60**	Text

Table 1: Quantitative comparison with baselines for **in-domain/out-of-the-domain** settings. The bold values indicate the best results, while the underlined values represent the second-best. Guid. indicates the modality of emotional guidance. Since there is no ground truth video in the out-of-the-domain setting, the FID metric is left empty. It can be observed that our model outperforms the baselines across many metrics.

input includes audio and emotional guidance (such as labels, driving videos, and text, depending on the method). For facial motion control, we use only textual instructions to drive the avatar.

4.1 Experimental Setups

Datasets and Implementation Details For emotional talking control, we augment the MEAD dataset (Wang et al. 2020) following the methods outlined in Sec. 3.2. MEAD is a large-scale emotional talking face dataset featuring 8 emotion types and 3 intensity levels. We reserved 5 individuals for testing purposes and utilized the remaining data for training. For text-guided facial motion control, we leveraged the CC v1 dataset (Hazirbas et al. 2021), which offers paired data comprising instructions and corresponding action videos. To ensure effective lip synchronization, we also incorporated the HDTF dataset (Zhang et al. 2021), which has high-quality talking face recordings. Our model was trained on a combination of these three datasets. The evaluation was conducted using MEAD for in-domain assessment and TalkingHead 1KH (Wang, Mallya, and Liu 2021b) for out-of-domain evaluation. We use Conformer (Gulati et al. 2020) as the backbone of our diffusion-based motion generator. Specifically, the model comprises 12 Conformer blocks, with a hidden state size of 768. For encoding textual instructions, we apply CLIP-L/14 (Radford et al. 2021), and the Adapters are two layers MLPs. We adopt the Adam (Kingma and Ba 2014) optimizer and train our models on 8 V100 GPUs.

Evaluation Metrics To measure the fine-grained controlling ability, we propose AU_{F1}, which calculates the F1 score of action units between the generated results and the ground truth. Furthermore, we introduce AU_{Emo}, calculated as how many action units could be recalled by typical AUs of specific emotion types, to evaluate the overall coverage of facial details w.r.t an emotion type in the generated video. For motion control, we introduce the $CLIP_S$ metric, which computes the CLIP embedding similarity between the text instruction and each frame, with the maximum value indicating the correspondence between the instruction and generated motion. Moreover, $Sync_D$ is utilized to gauge lip-sync quality using SyncNet (Chung and Zisserman 2017), and FID (Heusel et al. 2017) for both emotion and motion control to evaluate overall quality.

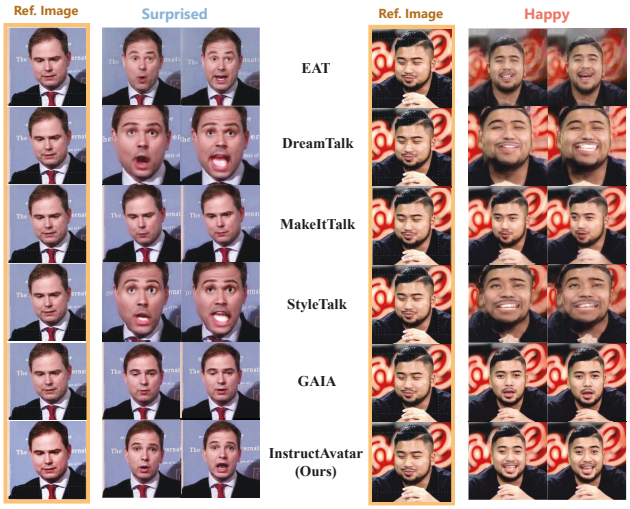

Figure 5: Qualitative comparison with baselines. It shows that InstructAvatar achieves good lip-sync quality and emotion controllability. Additionally, the outputs generated by our model exhibit enhanced naturalness and effectively preserve identity characteristics.

For subjective evaluation, we involve 15 experienced users to score the generation quality and controllability of each model. For emotional talking control, we assess the lip-sync quality (Lip.), emotion controllability (Emo.), naturalness (Nat.), and motion jittering (Jit.). For facial motion control, we measured the accuracy of instruction following (Mot.) and identity preservation (ID.). Participants were presented with one video at a time and asked to rate each video for each score on a scale of 1 to 5. We calculated the average score as the final result.

4.2 Experimental Results

Emotional Talking Control We compare InstructAvatar with several state-of-the-art methods. We consider emotion-unaware models GAIA (He et al. 2024) and MakeItTalk (Zhou et al. 2020), and provide the portrait and audio from the MEAD test set in inference. For the emotional label-based method EAT (Gan et al. 2023), we supplement the model with additional ground truth emotion types obtained in the annotation. Reference video-based methods like StyleTalk (Ma et al. 2023b) and DreamTalk (Ma et al.

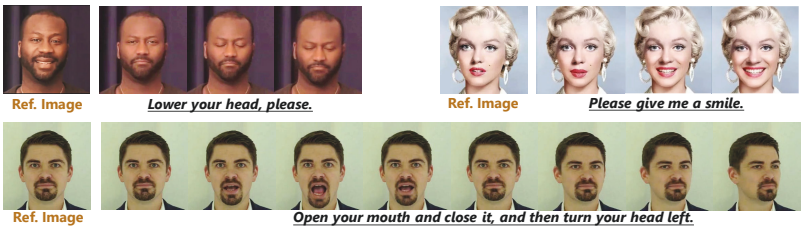

Figure 6: Examples of text-guided facial motion control. Our model can execute precise motion control and generalize effectively to connect different actions.

Methods	Obj. Met.		Subj. Met.	
	FID↓	CLIP_S↑	Mot.↑	ID.↑
Ref.	–	20.500	0	–
Rand. Inst.	27.782	21.274	0.19	4.85
GT Inst.	25.370	23.043	4.65	4.86
GT Video	–	23.136	5	5

Table 2: Objective and subjective metrics for text-guided facial motion control.

Methods	Sync_D↓	AU_F1↑	Mot.↑
InstructAvatar	9.653	0.552	4.52
(a) w/o AU	9.843	0.435	–
(b) w/o Branch	9.994	0.488	4.10
(c) w/o Zero conv.	12.832	0.434	3.62
(d) w/o Rand. sub.	9.597	0.339	–

Table 3: Ablation studies on the proposed techniques.

2023c) utilize the ground truth emotional video in MEAD as style guidance.

As shown in Tab. 1, the proposed InstructAvatar model demonstrates strong performance across both objective and subjective metrics in both in-domain and out-of-domain settings. Notably, our model achieves superior emotion controllability compared to baselines tailored for emotional talking, such as EAT, as evidenced by AU_F1 and Emo. scores. Furthermore, our model exhibits excellent lip-sync ability, surpassing GAIA, a recent state-of-the-art audio-driven talking head model, in the in-domain setting. Our Sync_D metric is closer to the ground truth video (9.172) and shows better FID scores. More importantly, the generated results from our model appear more natural and robust to portrait images, as reflected by the Nat. metric and the examples shown in Fig. 5. It is also worth noting that our model infers talking emotion solely from text inputs, which poses a more challenging task. Additionally, our model supports a broader range of instructions beyond high-level emotion types, a feature absent in most baselines.

Facial Motion Control To evaluate the effectiveness of the motion controllability, we establish four evaluation settings: **(i)** Simply repeating the portrait image (Ref.), **(ii)** Given a random instruction (Rand. Inst.), **(iii)** Given true instruction (GT Inst.), and **(iv)** Ground-truth video (GT video). We set up **(i)** as a baseline to assess whether our model can generate dynamic video, and **(ii)** to evaluate whether our model can follow instructions. Results for different metrics in these settings are presented in Tab. 2, and visual results are provided in Fig. 6. Note that the GT Inst. setting represents the typical inference manner. Our model exhibits accurate instruction-following ability, as evidenced by similar CLIP_S metrics with the ground-truth video and a large performance gap with Rand. Inst., along with high subjective motion accuracy (Mot.) scores. Moreover, our model demonstrates excellent video generation quality, producing natural portraits while maintaining identity, as indicated by the ID. metric and Fig. 6. Additionally, it generalizes effectively to connect different actions, as shown in Fig. 6.

4.3 Ablation Study

We conduct ablation studies to verify the effectiveness of each component of our model, as presented in Tab. 3. We utilize Sync_D to reflect lip-sync quality, AU_F1 to gauge emotion controllability, and Mot. to assess motion controllability. We can see that **(a)** In terms of data format, when no action units are provided during training, the model loses its ability to capture fine-grained details, resulting in a decrease in AU_F1. **(b)** Combining emotion learning and motion learning into a single branch leads to a contraction to some extent, negatively impacting the performance of both controls, as evidenced by the AU_F1 and Mot. metrics. **(c)** The Zero Convolution mechanism, detailed in Sec. 3.3, is designed to stabilize training and harvest abundant knowledge from pretrained talking face models. Removing this component dramatically deteriorates lip-sync quality, and also influences emotion and motion control. **(d)** To prevent emotion leakage, we substitute the key frame latent with another emotion type during training emotional talking data. We found that this method significantly improves emotion control ability during out-of-domain tests, where portraits typically exhibit neutral emotion.

5 Conclusion

In this paper, we introduce InstructAvatar, a novel text-guided unified framework for emotional talking and facial motion control in avatar generation, significantly enhancing controllability and vividness compared to previous models. Experimental results demonstrate InstructAvatar's exceptional lip-sync quality, fine-grained emotion controllability, user-friendly control interface, and naturalness of the generated outputs. We hope our work will inspire further research into text-guided 2D emotional talking heads and anticipate more studies in this area.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant No. 92470205.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Cao, H.; Cooper, D. G.; Keutmann, M. K.; Gur, R. C.; Nenkova, A.; and Verma, R. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4): 377–390.

Chung, J. S.; and Zisserman, A. 2017. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, 251–263. Springer.

Defossez, A.; Synnaeve, G.; and Adi, Y. 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Feng, G.; Cheng, H.; Li, Y.; Ma, Z.; Li, C.; Qian, Z.; Miao, Q.; and Pun, C.-M. 2024. EmoSpeaker: One-shot Fine-grained Emotion-Controlled Talking Face Generation. arXiv:2402.01422.

Gan, Y.; Yang, Z.; Yue, X.; Sun, L.; and Yang, Y. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22634–22645.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Hazirbas, C.; Bitton, J.; Dolhansky, B.; Pan, J.; Gordo, A.; and Ferrer, C. C. 2021. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3): 324–332.

He, T.; Guo, J.; Yu, R.; Wang, Y.; Zhu, J.; An, K.; Li, L.; Tan, X.; Wang, C.; Hu, H.; et al. 2024. GAIA: Zero-shot Talking Avatar Generation. *ICLR*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liang, J.; and Lu, F. 2024. Emotional Conversation: Empowering Talking Faces with Cohesive Expression, Gaze and Pose Generation. arXiv:2406.07895.

Liu, C.; Lin, Q.; Zeng, Z.; and Pan, Y. 2024. EmoFace: Audio-driven Emotional 3D Face Animation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE.

Liu, X.; Wu, Q.; Zhou, H.; Du, Y.; Wu, W.; Lin, D.; and Liu, Z. 2022. Audio-Driven Co-Speech Gesture Video Generation. *Advances in Neural Information Processing Systems*, 35: 21386–21399.

Luo, C.; Song, S.; Xie, W.; Shen, L.; and Gunes, H. 2022. Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI-2022. International Joint Conferences on Artificial Intelligence Organization.

Ma, Y.; Wang, S.; Ding, Y.; Ma, B.; Lv, T.; Fan, C.; Hu, Z.; Deng, Z.; and Yu, X. 2023a. TalkCLIP: Talking Head Generation with Text-Guided Expressive Speaking Styles. *arXiv preprint arXiv:2304.00334*.

Ma, Y.; Wang, S.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; Deng, Z.; and Yu, X. 2023b. Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv preprint arXiv:2301.01081*.

Ma, Y.; Zhang, S.; Wang, J.; Wang, X.; Zhang, Y.; and Deng, Z. 2023c. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*.

OpenAI. 2023. GPT-4V(ision) System Card.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, X.; Zhang, L.; Zhu, H.; Zhang, P.; Zhang, B.; Ji, X.; Zhou, K.; Gao, D.; Bo, L.; and Cao, X. 2023. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. *arXiv preprint arXiv:2312.01841*.

Sun, Y.; Chu, W.; Zhou, H.; Wang, K.; and Koike, H. 2024. AVI-Talking: Learning Audio-Visual Instructions for Expressive 3D Talking Face Generation. arXiv:2402.16124.

Tan, S.; Ji, B.; Bi, M.; and Pan, Y. 2024. EDTalk: Efficient Disentanglement for Emotional Talking Head Synthesis. arXiv:2404.01647.

Tan, S.; Ji, B.; and Pan, Y. 2023. EMMN: Emotional Motion Memory Network for Audio-driven Emotional Talking Face Generation. 22089–22099.

Tan, S.; Ji, B.; and Pan, Y. 2024. Style2Talker: High-Resolution Talking Head Generation with Emotion Style and Art Style. arXiv:2403.06365.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. arXiv:2402.17485.

Tripathy, S.; Kannala, J.; and Rahtu, E. 2021. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1329–1338.

Wang, C.; Tian, K.; Zhang, J.; Guan, Y.; Luo, F.; Shen, F.; Jiang, Z.; Gu, Q.; Han, X.; and Yang, W. 2024. V-Express: Conditional Dropout for Progressive Training of Portrait Video Generation. arXiv:2406.02511.

Wang, D.; Dai, B.; Deng, Y.; and Wang, B. 2023a. AgentAvatar: Disentangling Planning, Driving and Rendering for Photorealistic Avatar Agents. arXiv:2311.17465.

Wang, D.; Deng, Y.; Yin, Z.; Shum, H.-Y.; and Wang, B. 2023b. Progressive Disentangled Representation Learning for Fine-Grained Controllable Talking Head Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17979–17989.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, 700–717. Springer.

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

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021a. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021b. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Xu, S.; Chen, G.; Guo, Y.-X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; and Guo, B. 2024. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. arXiv:2404.10667.

Zhai, S.; Liu, M.; Li, Y.; Gao, Z.; Zhu, L.; and Nie, L. 2023. Talking Face Generation With Audio-Deduced Emotional Landmarks. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, B.; Qi, C.; Zhang, P.; Zhang, B.; Wu, H.; Chen, D.; Chen, Q.; Wang, Y.; and Wen, F. 2022. MetaPortrait: Identity-Preserving Talking Head Generation with Fast Personalized Adaptation. *arXiv:2212.08062*.

Zhang, C.; Wang, C.; Zhang, J.; Xu, H.; Song, G.; Xie, Y.; Luo, L.; Tian, Y.; Guo, X.; and Feng, J. 2023a. DREAM-Talk: Diffusion-based Realistic Emotional Audio-driven Method for Single Image Talking Face Generation. *arXiv preprint arXiv:2312.13578*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023b. 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*, 8652–8661.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhao, Q.; Long, P.; Zhang, Q.; Qin, D.; Liang, H.; Zhang, L.; Zhang, Y.; Yu, J.; and Xu, L. 2024. Media2face: Co-speech facial animation generation with multi-modality guidance. *arXiv preprint arXiv:2401.15687*.

Zhong, Y.; Wei, H.; Yang, P.; and Wang, Z. 2023. ExpCLIP: Bridging Text and Facial Expressions via Semantic Alignment. *arXiv preprint arXiv:2308.14448*.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6): 1–15.