InstructAvatar: Text-Guided Emotion and Motion Control for Avatar Generation (Supplementary Materials)

Anonymous submission

1 More Experimental Results

1.1 More Results about Emotional Talking Control

Additional outcomes concerning text-guided emotional talking control are presented in Fig. 1. We can see that our model exhibits precise emotion control ability, with the generated results appearing natural. Furthermore, InstructAvatar supports fine-grained control and demonstrates reasonable generalization ability beyond the domain.

1.2 More Results about Facial Motion Control

We show more results about facial motion control in Fig. 2. It is evident that InstructAvatar exhibits remarkable proficiency in following instructions and preserving identity. Furthermore, the generated results appear natural and robust with variations in the provided portrait, including tilting or inherent expressions. Moreover, our model demonstrates fine-grained control capability and performs effectively in out-of-domain scenarios, as depicted in the last row of Fig. 2.

1.3 The Effectiveness of Textual Instructions

To animate the avatar in our model, we input three conditions: portrait, audio, and textual instructions. We acknowledge that in real life, all these conditions can convey emotion. Therefore, a natural question arises: Does the emotion depicted in our generated videos primarily stem from the textual instructions rather than from the portraits or the inherent emotion conveyed in the audio? We address this question in Fig. 3, where all videos are generated using identical neutral portraits and neutral audio. The results demonstrate that our model can still produce distinct emotional talking videos, highlighting the effectiveness of textual instructions.

1.4 Emotion Intensity

InstructAvatar demonstrates the capability to generate results with varying levels of emotion intensity. We illustrate a case in Fig. 4. It is evident that our model distinguishes between different emotion intensities based on specific descriptors such as “extremely” and “slightly”.

Table 1: More ablation studies on the proposed techniques.

Methods	$\text{Sync}_{\mathbf{D}}\downarrow$	$\text{AU}_{\mathbf{F1}}\uparrow$	Mot.↑
InstructAvatar	9.747	0.537	4.46
(a) w/o CLIP Adapter	9.903	0.512	4.31
(b) w/o AU loss	9.753	0.504	–
(c) w/o Empty noise	10.427	0.469	3.23
(d) w/ Label input	9.763	0.469	–
(e) All tokens for Emo.	9.842	0.417	–
(f) `[EOS]` token for Mot.	–	–	4.03

1.5 More Ablation Results

We provide more ablation results in Tab. 1. To reduce the training cost, we randomly select 40% of the samples from the training pool. Given the substantial number of samples (over 60k), the statistical distribution differences would be marginal, thus representing the entire dataset. We observe that: **(a)** When the CLIP Adapter (Gao et al. 2024) is removed, the model loses a converter from the CLIP text space to the space required by the denoising block. This results in a decrease in both $\text{AU}_{\mathbf{F1}}$ and Mot. metrics. **(b)** When the AU loss is removed, the model loses some strict guidance on capturing fine-grained action unit details, leading to a decrease in $\text{AU}_{\mathbf{F1}}$. **(c)** To integrate facial motion control where no audio is provided into our unified framework, we use pseudo-empty audio as a placeholder. However, upon switching this strategy to employ another pseudo audio feature, such as tensors with all 0s, we observe a rapid deterioration in facial motion control, which also impacts emotional control, as indicated by Mot. and $\text{AU}_{\mathbf{F1}}$. We attribute this to the inherent physical meaning carried by pseudo-empty audio, symbolizing silence and resulting in a still avatar. Conversely, the use of fake audio features like tensors with all 0s lacks meaningful interpretation. Consequently, when combined with normal audio during training, the model becomes confused due to this misaligned setting. **(d)** Beyond the limited emotion categories used in previous models, we adopt natural language for an open-vocabulary emotion guidance approach. This method has several advantages as discussed in Sec.1 in the main text: (1) Enhanced control over fine-grained details rather than just the overall style; (2) Improved generalizability compared to limited emotion cate-

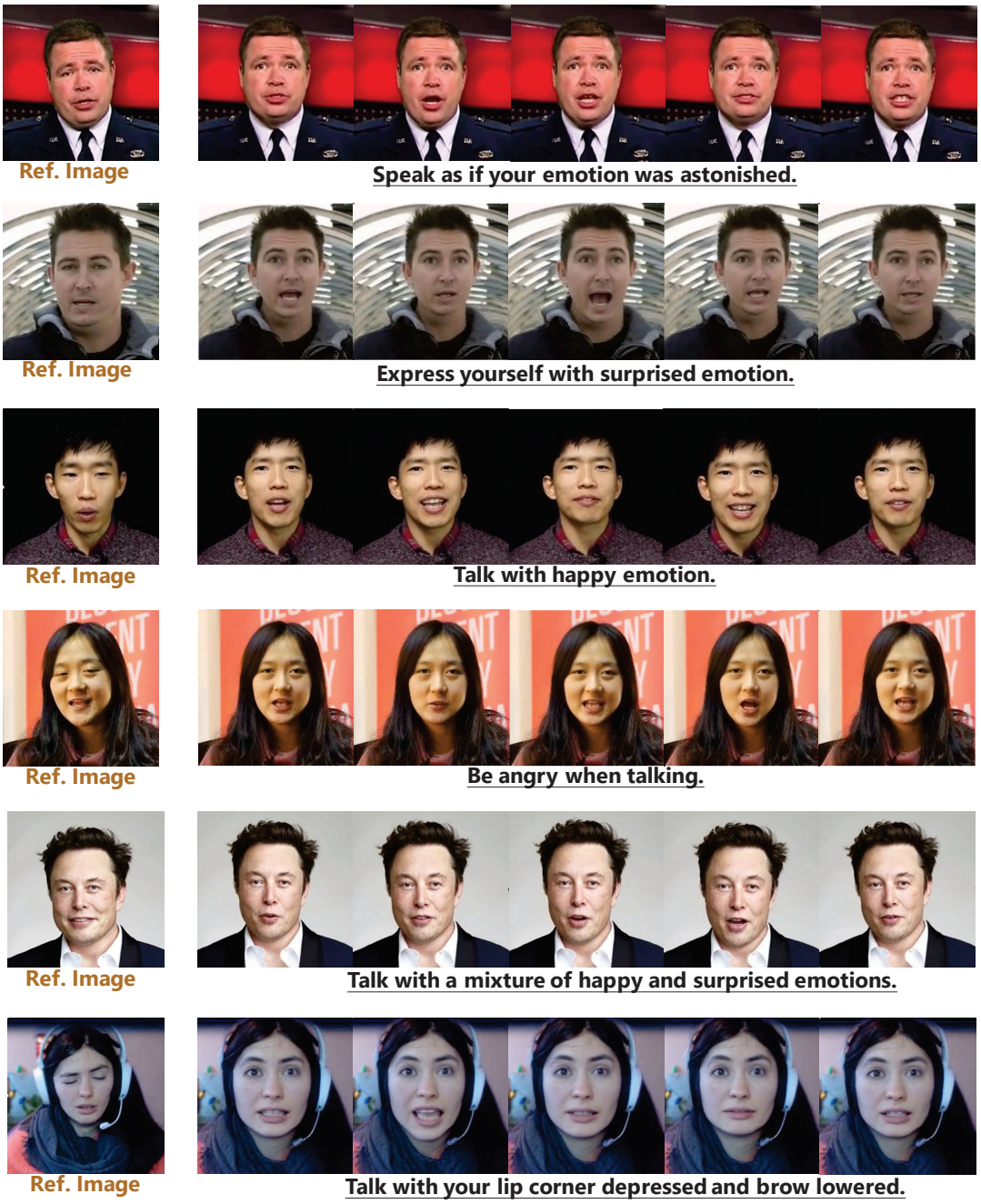

Figure 1: More examples of text-guided emotional talking control.

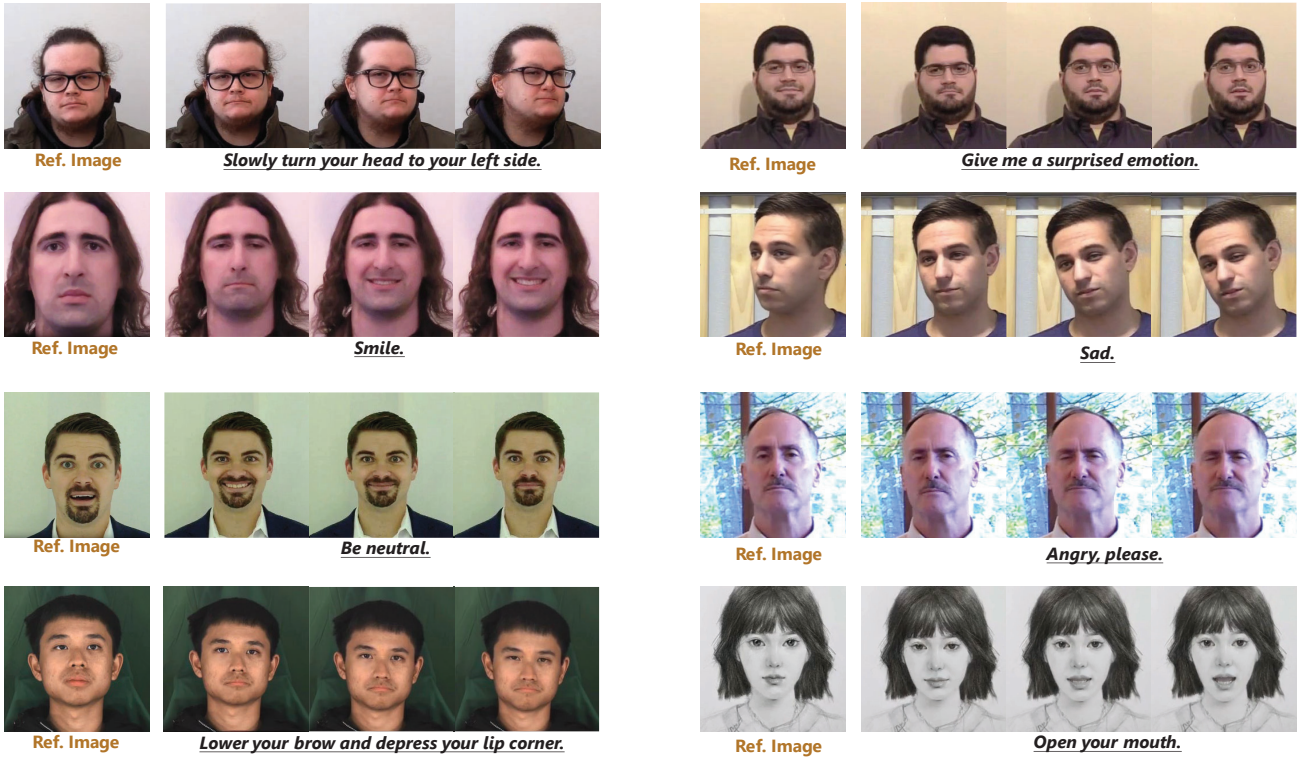

Figure 2: More examples of text-guided facial motion control.

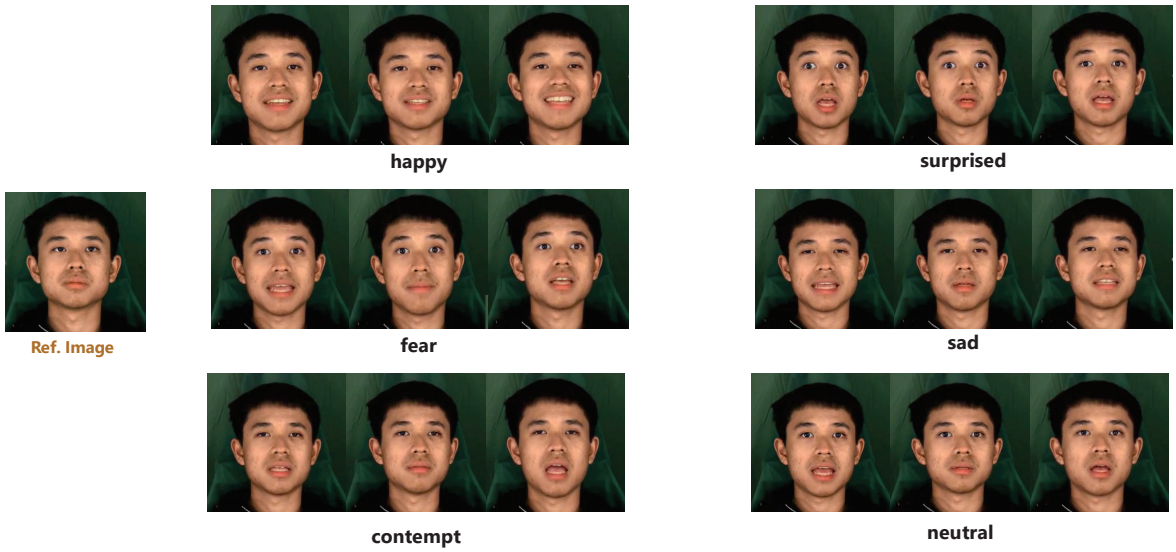

Figure 3: Illustration of the effectiveness of textual instructions. All videos are generated utilizing identical portraits and neutral audio, with variations only in the textual instructions.

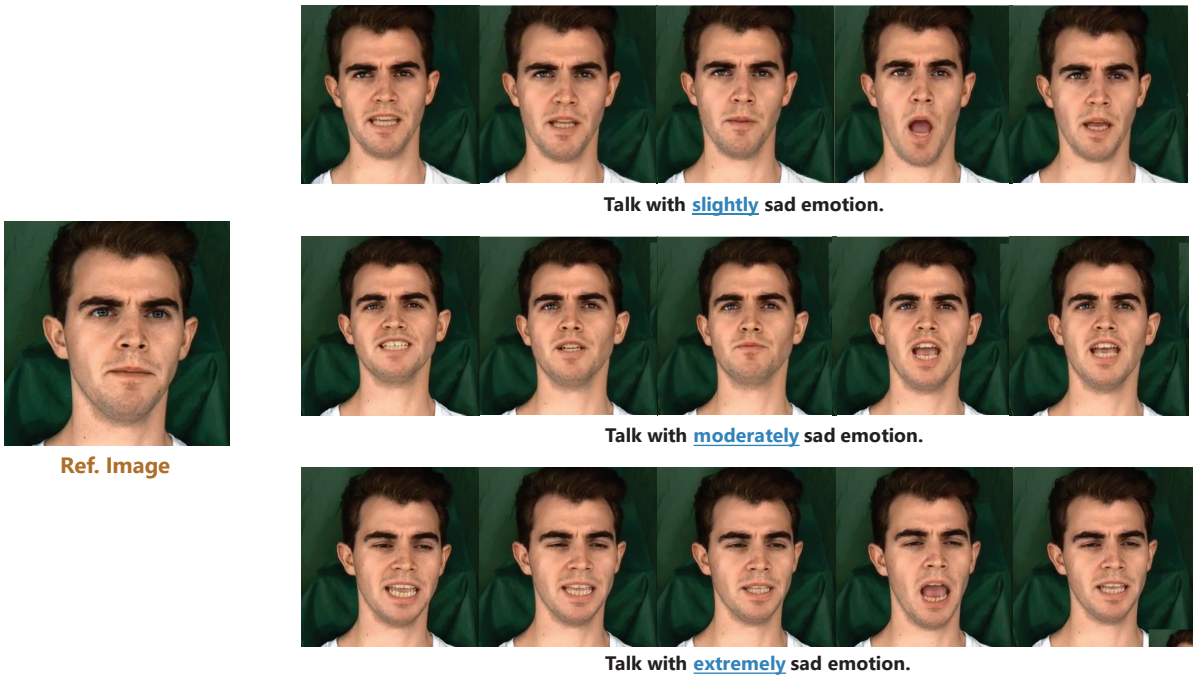

Figure 4: Illustration of emotion intensity control.

gories; and (3) Enhanced interactivity and user-friendliness. Experimentally, when training our model with emotion labels, as shown in Tab. 1 (d), the analysis of $\text{AU}_{\textbf{F1}}$ indicates that text guidance offers better fine-grained controllability. **(e)** The `[EOS]` token is well-suited for the emotional talking task, acting as a general style guide. Replacing it with all instruction tokens leads to a performance drop, demonstrating that additional tokens may distract the model from extracting overall information. **(f)** Conversely, the hidden states of all tokens are more suitable for the motion control task, which requires temporally dynamic guidance. Replacing these with only `[EOS]` tokens fails to reflect dynamic information, resulting in suboptimal performance.

2 Metrics Definition

We provide formal definitions for self-defined metrics in our paper.

2.1 AU_{F1} and AU_{Emo}

To evaluate the fine-grained controllability of InstructAvatar, we introduce $\text{AU}_{\textbf{F1}}$ and $\text{AU}_{\textbf{Emo}}$. Let us consider action units extracted from generated sample j, denoted by $\hat{\boldsymbol{y}}^{(j)} \in \mathbb{R}^M$, alongside corresponding ground truth action units $\boldsymbol{y}^{(j)} \in \mathbb{R}^M$, where M represents the number of action units (in our paper $M = 41$). Both $\hat{\boldsymbol{y}}^{(j)}$ and $\boldsymbol{y}^{(j)}$ are vectors composed of 0s and 1s, with 1 indicating activation of the action unit and 0 indicating its inactivation.

To quantify the concordance between the action units of generated results and ground truth, assuming we have n samples, we compute the F1 score for this multi-label classification problem as follows:

$$\text{AU}_{\textbf{F1}} = \frac{1}{n}\sum_{j=1}^{n}\frac{2|\boldsymbol{y}^{(j)} \cap \hat{\boldsymbol{y}}^{(j)}|}{|\boldsymbol{y}^{(j)}| + |\hat{\boldsymbol{y}}^{(j)}|} = \frac{1}{n}\sum_{j=1}^{n}\frac{2\sum_{i=1}^{M}\boldsymbol{y}_i^{(j)} \cdot \hat{\boldsymbol{y}}_i^{(j)}}{\sum_{i=1}^{M}\boldsymbol{y}_i^{(j)} + \sum_{i=1}^{M}\hat{\boldsymbol{y}}_i^{(j)}}$$

Moreover, to evaluate the overall coverage of facial details with respect to an emotion type, we define $\text{AU}_{\textbf{Emo}}$. Firstly, we identify typical and representative action unit combinations for each emotion, as shown in Tab. 2. Then, for each generated result $\hat{\boldsymbol{y}}^{(j)}$, suppose the corresponding action units for such desired emotion are $\boldsymbol{y}_{emo}^{(j)}$, we calculate how many action units could be recalled by the typical action units:

$$\text{AU}_{\textbf{Emo}} = \frac{1}{n}\sum_{j=1}^{n}\frac{2|\boldsymbol{y}_{emo}^{(j)} \cap \hat{\boldsymbol{y}}^{(j)}|}{|\boldsymbol{y}_{emo}^{(j)}|} = \frac{1}{n}\sum_{j=1}^{n}\frac{2\sum_{i=1}^{M}\boldsymbol{y}_{emo,i}^{(j)} \cdot \hat{\boldsymbol{y}}_i^{(j)}}{\sum_{i=1}^{M}\boldsymbol{y}_{emo,i}^{(j)}}$$

2.2 CLIP_{S}

We employ CLIP_{S} to measure the accuracy of text-guided motion control, as shown in Sec.4.1 in the main text. Let t denote the text instruction and v denote the generated video. We denote v_i as the i-th frame of video v. We use the CLIP (Radford et al. 2021) text encoder $\mathcal{E}_t$ to encode the text and the image encoder $\mathcal{E}_v$ to encode each frame. Since the CLIP model is trained on paired image-text data, it possesses powerful modality alignment ability. We utilize cosine similarity to calculate the matchness of each frame with the instruction. Considering that motion control is a dynamic process, we consider it successful if it matches the instructions for a period of time. Therefore, we take the maximum of the similarity scores s_i as our final result:

$$s = \max_i s_i = \max_i \frac{\mathcal{E}_t(t) \cdot \mathcal{E}_v(v_i)}{\|\mathcal{E}_t(t)\| \cdot \|\mathcal{E}_v(v_i)\|}$$

Table 2: Typical Action Units for Different Emotions.

Emotion Type	Typical Action Units
Angry	Brow Lowerer, Jaw Drop, Nose Wrinkler, Lid Tightener
Fear	Inner Brow Raiser, Jaw Drop, Upper Lid Raiser, Outer Brow Raiser
Happy	Cheek Raiser, Lip Corner Puller, Jaw Drop, Lid Tightener
Contempt	Inner Brow Raiser, Chin Raiser, Lip Corner Puller, Lid Tightener
Disgusted	Brow Lowerer, Cheek Raiser, Lip Corner Depressor, Nose Wrinkler
Sad	Brow Lowerer, Chin Raiser, Inner Brow Raiser, Lip Corner Depressor
Surprised	Inner Brow Raiser, Jaw Drop, Outer Brow Raiser, Lid Tightener

3 Data

3.1 Modalities of Each Dataset

In Tab. 3, We list the available modalities and corresponding tasks of each dataset utilized in our paper.

3.2 Data Preprocessing

We preprocess all datasets and establish filter policies to discard low-quality samples.

For video data, we standardize each clip into a portrait-centered talking head video with dimensions of 256 by 256 pixels and a frame rate of 25 fps. Our filtering strategy follows (He et al. 2024) summarized as follows: **(1)** We maintain consistency in the orientation of individuals facing the camera throughout video clips. Frames exhibiting significant deviations, potentially obscuring lip movements, are excluded. **(2)** We monitor the positions of faces across frames, ensuring minimal displacement over consecutive timestamps to achieve smooth facial motion in video clips. **(3)** Frames featuring individuals wearing masks or remaining silent are identified and removed. Additionally, to minimize domain gaps across datasets, we estimate the distribution of talking head position and scale in the HDTF (Zhang et al. 2021) videos and adjust the other datasets accordingly to this standard.

For each video clip obtained, we extract the audio and resample it to a 16kHz sampling rate. We normalize the speech and apply a denoiser (Defossez, Synnaeve, and Adi 2020) to reduce background noise. In the CC v1 dataset (Hazirbas et al. 2021), where the audio comprises off-screen instructional speech unrelated to lip movement, we generate pseudo-empty audio with zero amplitude and a duration matching that of the corresponding ground truth video clip. Subsequently, we extract audio features using Wave2Vec 2.0 (Baevski et al. 2020).

An important aspect of our model is its integration of textual information as a supervised signal. For emotional talking control, we outline the process of constructing textual instructions in Sec.3.2 in the main text, along with listing the templates and prompts used in Sec. 3.4. For facial motion control, the CC v1 dataset provides annotations consisting of off-screen instructional speeches obtained via ASR(Automatic Speech Recognition), along with corresponding timestamps. We extract the instructional annotations and prompt GPT-4V (OpenAI 2023) to paraphrase them into fluent sentences, eliminating incomplete forms due to ASR detection. Based on these timestamps, we extract the corresponding action videos. Regarding the HDTF dataset, which lacks explicit instructions, and considering that the majority of videos in this dataset exhibit neutral emotions, we provide pseudo instructions such as "Talk with neutral emotion" or "Talk with an emotionless face". etc.

3.3 Data Statistic

We provide the statistics of each dataset in Tab. 4.

3.4 Instruction Templates and GPT-4V Prompt

We present a portion of the templates utilized for transforming emotion types into sentences in Tab. 5, along with the prompts used to query GPT-4V in Tab. 6.

4 Implementation Details

4.1 Illustration of VAE

In our model, we employ a Variational Autoencoder (VAE) based on the framework outlined in (He et al. 2024). The VAE is designed to disentangle motion information from video data and consists of two encoders: the motion encoder and the appearance encoder, along with a single decoder.

To prevent the leakage of appearance information in reconstruction, they utilize the appearance information from the i-th frame and the motion information from the j-th frame to reconstruct the j-th frame by the VAE. Therefore, in cases where the i-th and j-th frames from one video clip contain the same appearance but different motion information (e.g., the same person speaking different words), the VAE model learns to first extract the pure appearance feature from the i-th frame. Subsequently, it combines this feature with the pure motion feature of the j-th frame to accurately reconstruct the original j-th frame.

Once the VAE is well trained, it can encode a video into disentangled appearance and motion latents. For a video with l frames, we can randomly select a frame for appearance encoding, resulting in an appearance latent of shape $(1, d_{app})$, where d_{app} is 768 for the $\text{GAIA}_{\text{base}}$ model. This corresponds to the flattened dimensions $(3, 16, 16)$, capturing spatial information. For the motion latent, we obtain a

Table 3: Available modalities and corresponding tasks of the collected dataset. * indicates that for TalkingHead 1KH, we only use the portrait images to measure the out-of-domain generative ability of our model.

Dataset	Task	Video	Audio (Speech)	Instruction
MEAD	Emotinoal Talking Control	✓	✓	✓
HDTF		✓	✓	✓(Always neutral)
CC v1	Facial Motion Control	✓		✓
TalkingHead 1KH	Evaluation*	✓	✓	

Table 4: Statistics of the collected dataset.

Datasets	Raw		Filtered	
	#IDs	#Hours	#IDs	#Hours
HDTF	362	16	359	14
CC v1 (Motion Control)	2,412	29	2,412	23
MEAD	47	42	47	31
Total	2,821	87	2,818	68

latent of shape (l, d_{mot}) by encoding all frames with the motion encoder, where d_{mot} is also 768.

4.2 Hyperparameters and Model Architectures

The VAE comprises cascaded traditional convolutional residual blocks for both appearance and motion encoders, with downsampling factors of 8 and 16, respectively. The hidden size and number of layers are set to 256 and 4, resulting in approximately 700M parameters. The motion generator, based on diffusion models (Ho, Jain, and Abbeel 2020), consists of 12 Conformer (Gulati et al. 2020) blocks with a hidden state size of 768. The total number of parameters for the motion generator is 409M, including around 100 million for the non-trainable CLIP-L/14 (Radford et al. 2021) text encoder and approximately 300 million for the denoising backbone. The adapter (Gao et al. 2024) utilizes a two-layer MLP with skip connections, where the hidden state size of the middle layer is 4 units smaller than the input dimension.

During training, the Adam (Kingma and Ba 2014) optimizer is employed with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate starts at 1e-5 and follows an inverse square root schedule with 8000 warmup steps. For the diffusion model, a quadratic β schedule is set with $\beta_{min} = 0.05$ and $\beta_{max} = 20$. During inference, the model follows the DDIM (Song, Meng, and Ermon 2020) approach and samples 150 steps. The loss weights are set to $\lambda_{pose} = 1$, $\lambda_{au} = \lambda_{inten} = 0.1$.

5 Ethical Consideration and Limitation

For ethical considerations, InstructAvatar is designed to advance AI research on talking avatar generation. Responsible usage is strongly encouraged, and we discourage users from employing our model to generate intentionally deceptive content or engage in other inauthentic activities. To prevent misuse, adding watermarks is a common approach. Moreover, as a generative model, our results can be utilized to construct artificial datasets and train discriminative models.

Limitations Our work still has limitations. For example, our model is trained solely on a combination of action units extracted from real talking videos. This dependency between action units may limit its ability to precisely control a disentangled single action unit. Additionally, the relatively modest size of our training dataset may hinder its robustness when faced with highly out-of-domain instructions or appearances. Moreover, since almost all data in our training dataset follows a single emotion/motion pattern, it is challenging for our model to control both emotion and motion simultaneously. We leave these challenges for future exploration.

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.

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

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*.

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

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

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.

========== INSTRUCTION TEMPLATES ==========

”Talk with [EMO] emotion”,
”Please talk with a [EMO] expression.”,
”Please talk with a [EMO] face.”,
”Please talk with a [EMO] look.”,
”Please engage in a conversation using an [EMO] expression.”,
”Converse with the help of [EMO] expression.”,
”Make use of [EMO] expression as we discuss.”,
”Could you express [EMO]ly in your speech?”,
”speak with the expression of [EMO]”,
”Please convey your thoughts with the emotion of [EMO].”,
”I’d like you to talk as if you were feeling [EMO].”,
”Incorporate [EMO] into your speech.”,
”Let your words reflect the sentiment of [EMO].”,
”Let [EMO] be your guide in speaking.”,
”Express yourself as if you were experiencing [EMO].”,
”Speak as if your emotions were [EMO].”,
”Try conveying your message with [EMO] as your tone.”,
”Let [EMO] be your emotion as you speak.”,
”Speak with the emotion of [EMO] guiding your words.”,
”Let your words with the feeling of [EMO].”,
”I’d like you to communicate using [EMO] expression, please.”,
”Could you chat employing [EMO] expression?”,
”Discuss the topic while embracing the spirit of [EMO].”,
”Let’s discuss things while incorporating [EMO] expression.”,
”Feel free to use [EMO] expression while we talk.”,
”Speak as if you were surrounded by [EMO].”,
”Let the essence of [EMO] guide your words.”,
”Let [EMO] set the tone for your dialogue.”,
”Could you communicate with the spirit of [EMO]?”,
”Express your thoughts with the [EMO] in mind.”,
”Let [EMO] be the guiding principle in your conversation.”,
”Try conveying your message while being [EMO].”,
”Let [EMO] set the mood for your conversation.”,
”Speak as if your emotions were [EMO].”,
”Let [EMO] shape the texture of your conversation.”,
”Try conveying your message while being [EMO].”,
”Let [EMO] guides your dialogue.”,
”Speak as if [EMO].”,
”Express your ideas with the emotional of [EMO].”,
.....

Table 5: The templates used in transforming emotion types into a sentence.

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

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.

========== PROMPT TEMPLATES ==========

Action unit is a term used in facial expression analysis to describe specific movements of the facial muscles, which can be used to interpret and understand human emotions and expressions, For example, Inner Brow Raiser is the raising of the inner portion of the eyebrows, which is associated with surprise.

Now I have obtained action units from a video, the form is like this: ["brow_lowerer", "lips_part", "cheek_raiser"], I want to combine them into a sentence, like "make brow lower and separated lips, what's more, you can also lift your cheek".

More examples like this: ["upper_lip_raiser", "lips_part"] → " make sure lip raised and lips parted"; ["brow_lowerer", "lid_tightener"] → "drop brow, at the same time you can tighten your lid" ; ["lips_part", "lip_corner_puller"] → "try to part lips, meanwhile stretch lip corner".

Now please observe the image ***{img}*** above, I give predicted action units ***{au_list}***, please turn it into a natural and diverse sentence. If you find a contraction with the image, you can edit the action unit. Give me three examples.

Pay attention that the AU subject (like brow) should be maintained except for errors while the AU verb (like lower) can be changed. The way you express the sentence can also be free. You should try to make it diverse, clear, and natural but do not imagine an unrealistic subject. Avoid using "you" if possible. Do not use adverbs describing degree, such as "slightly". Do not incorporate temporal information, such as "begin", or "then".

Your answer:

Table 6: The prompts used to query GPT-4V.