Supplementary Material

Zhihao Xu[1*], Shengjie Gong[1*], Jiapeng Tang[2], Lingyu Liang[1], Yining Huang[1], Haojie Li[1], and Shuangping Huang[1,3†]

[1] South China University of Technology
[2] Technical University of Munich
[3] Pazhou Laboratory
{eezhihaoxu,eeshengjiegong}@mail.scut.edu.cn, eehsp@scut.edu.cn

In this supplementary material, we provide more implementation details on KMTalk (Sec. A), additional results and comparisons in Sec. B, and more discussion (Sec. C).

A Implementation Details

Network Architecture. To enhance the reproducibility of our KMTalk approach, we provide the detailed network architecture for Linguistic-based Key Motion Acquisition ((Sec. 3.2) and Cross-modal Motion Completion (Sec. 3.3) in the main paper. The network architecture is presented in Table 1. Our codebase will be released soon.

Phoneme-based Localization Method. The Phoneme-based Localization Method is proposed in this paper to locate the position of each phoneme. The specific procedure is as follows: First, the input speech signal is processed by an Automated Speech Recognition (ASR) module [4, 8], which transcribes the speech into its corresponding textual representation based on acoustic and language models. Subsequently, the Montreal Forced Aligner (MFA) [§] module is employed to establish temporal alignments between the transcribed text and the original speech signal. This module utilizes advanced algorithms to match the corresponding phonemes (in International Phonetic Alphabet format) within the transcribed text with their respective time locations in the speech waveform. Finally, the frame positions corresponding to the start and end timestamps of each phoneme are obtained, allowing for localization of the phoneme boundaries.

Details of Integration with Existing Methods We integrate pre-trained models of existing methods with phoneme-based localization techniques to construct different implementations of linguistic-based key motion capture. To elaborate, we generate a complete motion sequence using the pre-trained model of each existing method. Then, we extract key motions from the complete sequence based on the temporal position from phoneme-based localization. Subsequently,

[*] Authors contributed equally.
[†] Corresponding author.
[§] Montreal Forced Aligner (MFA): https://mfa-models.readthedocs.io/en/latest/mfa_phone_set.html

Table 1: Parameter illustration of network architectures. $\mathrm{L}(c_i, c_o)$ denotes a linear layer with input channels of c_i and output channels of c_o. $\mathrm{Concat}(v_1, v_2, c)$ stands for the concatenation of v_1 and v_2 in dimension c. Sigmoid represents a sigmoid function. Weighted $\mathrm{Sum}(W)$ denotes a weighted sum with the weight of W. $\mathrm{TransformerDecoder}(d\_model, nhead, dim\_ffd, num\_layers)$ represents a transformer structure with the input channels d_model, the number of heads in multi-head attention $nhead$, the channels of feedforward network dim_ffd and the number of decoder layers num_layers. $\mathrm{PE}(a)$ is a position embedding layer where a denotes the length of position vector. $\mathrm{MultiheadAttention}(d\_model, nhead)$ is an self-attention layer. $\mathrm{FFN}(d\_model)$ is a feed forward layer. Conv1D represents 1D convolution operation. The details of Manifold can be found in [5].

Module	Input → Output	Layer Operation
Audio Encoder	$\mathbf{x} \rightarrow \mathbf{A}(N, d)$	Wav2vec 2.0 pre-trained model [1]
Key Motion Decoder	$\mathbf{A_k}(m, d) \rightarrow \mathbf{K}(m, 3 \cdot V)$	$\mathrm{L}(d, f) \rightarrow \mathrm{TransformerDecoder}(f, 4, 2 \cdot f, 1) \rightarrow \mathrm{L}(f, 3 \cdot V)$
Motion Flow Encoder	$\mathbf{K}(m, 3 \cdot V) \rightarrow \mathbf{\Phi_k}(m, d)$	$\mathrm{PE}(16) \rightarrow \mathrm{L}(16{+}f, f) \rightarrow [\mathrm{MultiheadAttention}(f, 8) \rightarrow \mathrm{FFN}(f)]{\times}6$
	$\mathbf{T_k}(m) \rightarrow \mathbf{\Phi_{non\text{-}key}}(N-m, d)$	$\mathrm{PE}(16) \rightarrow \mathrm{L}(16, f) \rightarrow [\mathrm{MultiheadAttention}(f, 8) \rightarrow \mathrm{FFN}(f)]{\times}6$
	$\mathbf{\Phi_k}, \mathbf{\Phi_{non\text{-}key}} \rightarrow \mathbf{\Phi}(N, d)$	$\mathrm{Manifold}(\mathrm{FFN}(\mathbf{\Phi_k}), \mathbf{\Phi_{non\text{-}key}}) \rightarrow \mathrm{Conv1D} \rightarrow$ $[\mathrm{MultiheadAttention}(f, 8) \rightarrow \mathrm{FFN}(f)]{\times}6$ $\rightarrow \mathrm{Conv1D} \rightarrow \mathrm{L}(f, d)$
Motion Decoder	$\mathbf{A}(N, d), \mathbf{\Phi}(N, d) \rightarrow \mathbf{W}(N, d)$	$\mathrm{Concat}(\mathbf{A}, \mathbf{\Phi}, 2) \rightarrow \mathrm{L}(2 \cdot d, d) \rightarrow \mathrm{Sigmoid}$
	$\mathbf{A}(N, d), \mathbf{\Phi}(N, d) \rightarrow \mathbf{Z}(N, d)$	$\mathrm{Weighted\ Sum}(\mathbf{W})$
	$\mathbf{Z}(N, d) \rightarrow \mathbf{Y}(N, 3 \cdot V)$	$\mathrm{L}(d, f) \rightarrow \mathrm{TransformerDecoder}(f, 4, 2 \cdot f, 1) \rightarrow \mathrm{L}(f, 3 \cdot V)$

the CMC module is trained to extend the key motions from different methods into complete, continuous facial mesh sequences. For fair comparisons, the multi-modal motion decoder and loss calculation of the CMC module remain consistent with our method.

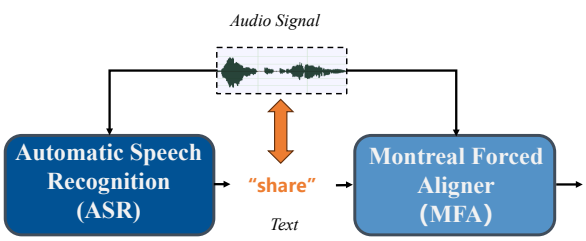
B Additional Results

B.1 Visualization Results of Phoneme Boundaries

To better comprehend the prior that articulatory actions are more pronounced at phoneme boundaries and effectively capture the kinematic characteristics of the entire motion sequence, we visualized the pronunciation of words alongside their corresponding lip offsets. We extracted audio fragments from the BIWI and VOCASET datasets, which are shown in Fig. 4. We observed that the key motion positions determined by the Phoneme-based Localization Method approximately capture the inflection points of the lip movement curve, denoted as key points. Once these key points are determined, the remaining frames can be effectively fitted using the linear interpolation method. Therefore, these key points can well describe the patterns of lip movement.

B.2 Integration with Existing Methods on VOCASET

The results of integrating our proposed progressive learning mechanism utilizing key motion embeddings with existing methods on VOCA-Test are shown in Table



Phoneme Time Position (i.e. share)

Phoneme	Start time(s)/frame	End time(s)/frame
/ʃ/	0.82(20)	0.93(23)
/ɛ/	0.93(23)	0.96(24)
/ə/	0.96(24)	1.05(27)
/r/	1.09(27)	1.11(28)

Fig. 1: The pipeline of the Phoneme-based Localization Method includes the Automatic Speech Recognition (ASR) module and the Montreal Forced Aligner(MFA) module.

Table 2: The results of integrating our proposed KMTalk with existing methods on VOCA-Test.

Methods		LVE↓ $\times 10^{-5}$mm	FDD↓ $\times 10^{-7}$mm
FaceFormer [3]	Original	4.1090	4.6675
	After	**3.9608**	**4.5343**
CodeTalker [10]	Original	3.9445	4.5422
	After	**3.8473**	**3.9043**
SelfTalk [7]	Original	3.2238	4.0912
	After	**2.6608**	**3.6795**

2. The experimental results demonstrate that our proposed learning mechanism can achieve significant improvements over existing state-of-the-art methods [3, 7, 10] on VOCASET, further confirming the strong generalization capabilities of our design.

B.3 Additional Results

Ablation Studies on VOCASET Ablation studies of KMTalk on VOCASET are presented in Table 3, and the results are consistent with the experiments conducted on BIWI. This further validates the effectiveness of the Phoneme-based Localization Method, Key Motion-focused Decoder, and Audio Guidance in CMC.

Ablation Studies of Loss Functions The latent consistency loss, measured by MSE, aligns latent audio features with lip encoder outputs, enhancing feature consistency. The text consistency loss, quantified by CTC, ensures lip movements match the source audio for accurate lip-reading. We empirically found that with the current weight strategy, the re-weighted losses are comparable, achieving the optimizal results. Ablation studies in Table 4 indicated a decrease in LVE without the use of these two losses, underscoring the importance of text and latent consistency loss for lip-reading accuracy.

Table 3: Ablation study for our components on VOCA-Test.

Phoneme-based Localization Method	Key Motion-focused Decoder	Audio Guidance in CMC	LVE↓	FDD↓
—	—	—	3.2238	4.0912
—	✓	✓	3.0987	4.1578
✓	—	✓	2.8402	4.0482
✓	✓	—	4.7366	5.2046
✓	✓	✓	**2.2639**	**4.0594**

Table 4: Additional ablations on BIWI-Test-A dataset.

Ablation	LVE↓ $\times 10^{-5}$mm	FDD↓ $\times 10^{-7}$mm
LKMA (wo/text loss and latent loss)	4.0604	**2.3137**
CMC (fusion with self-attention)	4.1824	3.3777
KMTalk(Ours)	**3.9654**	2.5446

Effectiveness of Fusion Module Design To validate the design of the CMC module, we conducted an ablation study that directly utilized a self-attention [2, 9] for multimodal fusion. The experimental results, presented in the third row of Table 4, indicate a decline in performance when using self-attention for multimodal fusion.

B.4 Results of Key Motions Quantity

The comparison results of key motion quantity are shown in Table 5. The results indicate that the quantity of key motions obtained with a uniform sampling stride of 3 is closest to the quantity obtained with the Phoneme-based Localization Method. Additionally, the experimental results suggest that uniform sampling does not consider the varying importance of different elements, and simply increasing or decreasing the quantity of key motions does not significantly improve the results. Therefore, proposing a prior to capture the varying importance of different elements is crucial for enhancing the model's performance.

B.5 Additional Quantitative Comparisons

Additional visual comparisons of facial meshes generated by various methods and ground truths are presented in Fig. 2. Our method consistently shows lower errors across diverse speech sequences, underscoring its proficiency in producing more accurate facial animations.

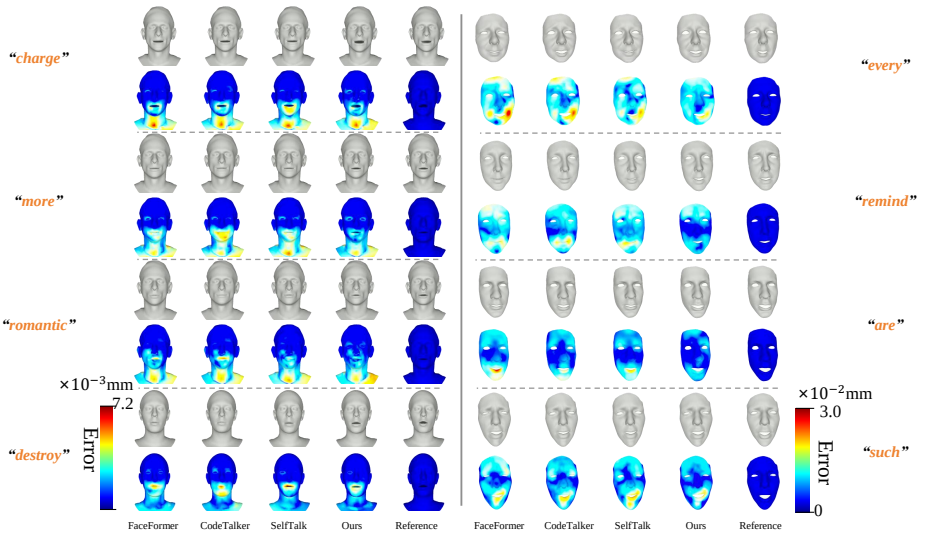


Fig. 2: Qualitative comparisons on VOCASET (left) and BIWI (right). We provide visual comparisons of facial animations synchronized with eight syllables extracted from the test speech sequences. The 1st, 3rd, 5th, and 7th rows display synthesized meshes and their corresponding ground-truths, while the 2nd, 4th, 6th, and 8th rows visualize the L2 loss for individual frames. Our method demonstrates more precise mouth movement and generates more natural and synchronized motion sequences visually.

B.6 Visualization of Long Sequence Generation

Both the VOCASET and BIWI datasets feature single-sentence inputs, typically under 5 seconds. Although we follow prior works [3, 7, 10] in experimenting with sentence-level datasets, our method can naturally extend to long sequences. We evaluated audio sequences including pauses with a duration of 1.5 minutes, and visualized the initial 30 seconds of intermediate frames and lip vertex displacement in Fig. 3. Our results can still produce accurate 3D talking face animation.For much longer audios, we segmented sequences into several clips and performed model inference for each clip individually.

B.7 User Study

The user study interface, designed for this research, is depicted in Fig. 5. The anticipated completion time for the user study is estimated to be between 10 to 15 minutes, considering 24 pairs of videos, each lasting 5 seconds, and three repetitions of watching. To mitigate the influence of random selection, we exclude comparison results completed in less than two minutes. Each participant is presented with the user study interface, which includes 24 video pairs. Participants are instructed to evaluate the videos twice, answering the following questions for each pair: "Compare the lips of the two faces: which one is more in

Table 5: Comparison Results of Key Motions Quantity on BIWI-Test-A.

Method	Quantity	Proportion	LVE ↓ $\times 10^{-4}$mm	FDD ↓ $\times 10^{-5}$mm
Uniform Sampling (Step 2)	1944	50.1%	4.1605	3.0792
Uniform Sampling (Step 3)	1301	33.5%	4.1648	2.8713
Uniform Sampling (Step 4)	980	25.3%	4.1655	2.7521
Phoneme-based Localization Method	1262	32.5%	**3.9742**	**2.5973**

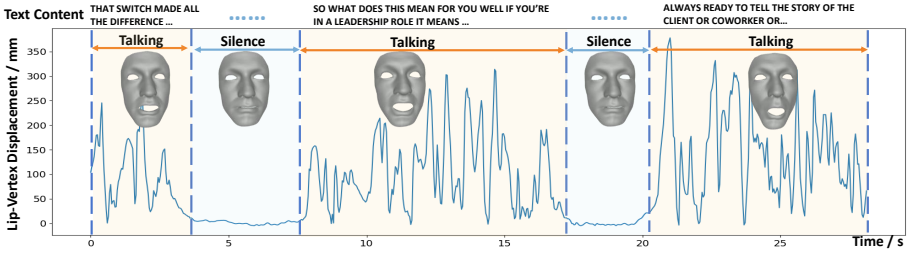


Fig. 3: Test the longer audio clips with pauses

sync (aligned) with the audio?" and "Compare the two full faces: which one is more realistic and trustworthy?". The user study interface facilitates the evaluation process, allowing participants to make informed judgments based on these specific criteria.

B.8 Video Comparison

To better evaluate the qualitative results produced by both our KMTalk and competing methods, we provide a supplementary video for demonstration and comparison. Specifically, we utilize a variety of audio clips to test our model, including segments extracted from TED videos, audio sequences from the VOCASET and BIWI datasets, as well as speech extracted from supplementary videos of previous methods. The video demonstrates the capability of KMTalk to synthesize facial animations with realistic and natural lip synchronization. It is worth noting that in comparison to competing methods such as FaceFormer [3], CodeTalker [10], and SelfTalk [7], which have experienced issues with over-smoothing, our KMTalk generates more dynamic and realistic facial movements with better lip synchronization. Furthermore, we demonstrate facial animations for speaking in different languages, such as Spanish, German, French, and more. The supplementary video serves as a visual demonstration, enabling a comprehensive comparison of the capabilities and strengths of our KMTalk approach. It highlights the ability of KMTalk to generate high-quality facial animations that exhibit natural lip movements, providing a more convincing and immersive user experience.

Table 6: Quantitative comparisons on VOCA-Test dataset.

Method	FPS	LVE↓ $\times 10^{-5}$mm	FDD↓ $\times 10^{-7}$mm
S2L+S2D	60	3.6467	4.0738
KMTalk	60	2.3115	4.0669
KMTalk	30	**2.2639**	**4.0594**

C Additional Discussions

C.1 Inference Time

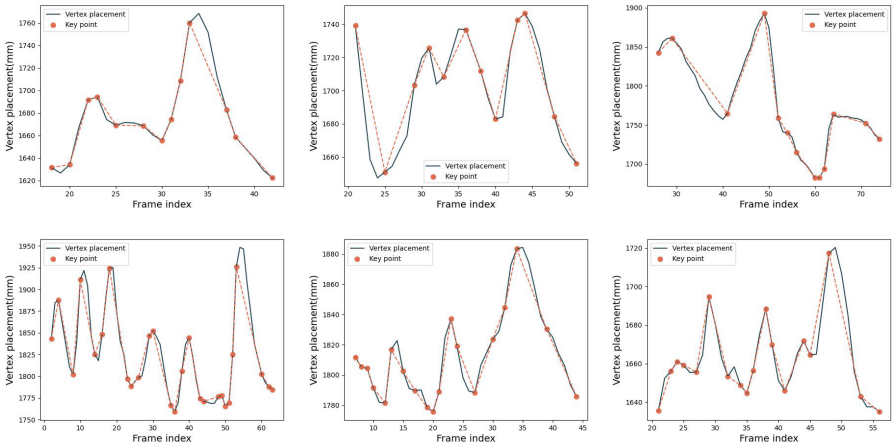
KMTalk's inference time on a single 3090 GPU for ASR [8] is 0.07 seconds and for MFA is 0.2 seconds on the BIWI dataset, with the LKMA and CMC modules together taking 0.37 seconds. Therefore, the average inference time for one audio clip is approximately 0.64 seconds. In comparison, Selftalk's inference time is 0.2 seconds per audio clip. Despite this increase, it is relatively minimal considering the complementary benefits of the modules and the overall performance of the system.

C.2 Frame Rate

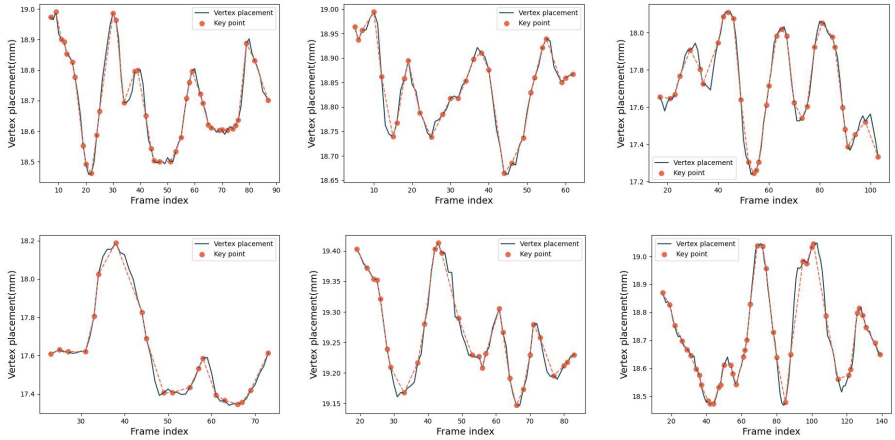
Our method, KMTalk, operates at a frame rate of 30 fps on VOCASET and 25 fps on BIWI, following the setting of previous methods [3,7,10]. The Audio analysis from our experiments indicates that phonemes have an average duration of approximately 0.1 seconds, thereby a frame rate exceeding 10 fps is adequate for identifying phoneme positions. Existing datasets, such as BIWI and VOCASET, with frame rates greater than 25 fps provide sufficient resolution to distinguish different phonemes. While a high frame rate (e.g., 60 fps) increases the number of frames between keyframes, potentially affecting the model's performance, our designed CMC module introduces global audio information, effectively mitigating the adverse effects of sparser keyframes. We compared S2L+S2D [6] in our setting and also adapted KMTalk to operate at 60 fps, and the results, shown in Table 6, demonstrate the superiority of KMTalk over S2L+S2D [6] on LVE and FDD metrics and confirm the robustness of our approach at higher frame rates.

C.3 Limitation Discussion

Our method requires the localization of keyframes, thus in challenging scenarios such as dialect variations, localization may involve standard keyframe detection errors. However, our method has demonstrated a certain degree of robustness even in the presence of deviation in keyframe localization. As shown in the last row of Table 4 in Sec. 4, our method still outperforms Selftalk by 26% in the FDD metric despite the presence of keyframe offset deviation. In the future, integrating advanced ASR technology could enhance the model's robustness and adaptability to various speech patterns.



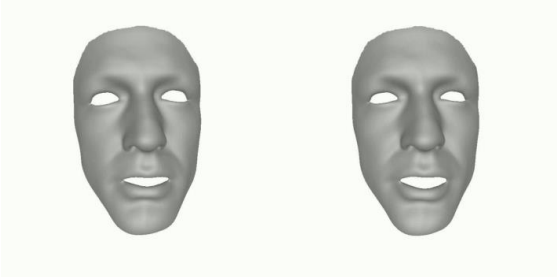
(a) Visualization results on BIWI.



(b) Visualization results on VOCASET.

Fig. 4: The visualization of phoneme boundaries on the BIWI and VOCASET datasets is presented separately in (a) and (b). Specifically, in this visualization, the vertex placement represents the cumulative Euclidean distance between the facial animation and the template in the lip region for each frame. The positions of key points are determined by the Phoneme Localization Method. Once these key points are marked, a linear interpolation method is employed to fit an approximate curve that closely approximates the marked key points.

Instructions:
Please watch 24 sets of provided videos(duration ~5s) of two facial animation. Carefully observe the full faces and lips, then choose a talking head (a or b) from the perspectives of synchronization and authenticity (one comparison, two questions) based on your observation. Please submit the questionnaire within 10-15 minutes.
Reminder:
For more efficient answering, please turn on the sound and use full screen playback on computer.
Comparison 1：



(a) (b)

1.1 Compare the lips of two faces, which one is more in sync (aligned) with the audio?

○ a

○ b

1.2 Compare the two full faces, which one is more realistic and trustworthy?

○ a

○ b

Fig. 5: Designed user study interface. Each participant need to answer 24 video pairs and here only one video pair is shown due to the page limit.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
2. Dai, G., Zhang, Y., Wang, Q., Du, Q., Yu, Z., Liu, Z., Huang, S.: Disentangling writer and character styles for handwriting generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5977–5986 (2023)
3. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: CVPR. pp. 18770–18780 (2022)
4. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M.: Auto-avsr: Audio-visual speech recognition with automatic labels. arxiv 2023. arXiv preprint arXiv:2303.14307
5. Mo, C.A., Hu, K., Long, C., Wang, Z.: Continuous intermediate token learning with implicit motion manifold for keyframe based motion interpolation. In: CVPR. pp. 13894–13903 (2023)
6. Nocentini, F., Ferrari, C., Berretti, S.: Learning landmarks motion from speech for speaker-agnostic 3d talking heads generation. arXiv preprint arXiv:2306.01415 (2023)
7. Peng, Z., Luo, Y., Shi, Y., Xu, H., Zhu, X., Liu, H., He, J., Fan, Z.: Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. arXiv preprint arXiv:2306.10799 (2023)
8. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023)
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
10. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: CVPR. pp. 12780–12790 (2023)