

3DiFACE: Synthesizing and Editing Holistic 3D Facial Animation

Supplementary Material

1. Additional Evaluation

Additional Holistic-motion synthesis evaluation: In addition to comparing against SadTalker [27] and Talkshow [26], we also conduct additional comparisons against DiffPoseTalk [22], a closely related concurrent work. To this end, we perform a qualitative and perceptual study to evaluate the holistic-motion synthesis and style-similarity. Specifically, we conducted multiple individual user studies in the form of A/B tests (see Section 2.7). As we see from Table 1, our method consistently outperforms all baselines in generating natural, realistic holistic motion with highly accurate lip synchronization. In the style-similarity studies presented in Table 2, our method matches the performance of Imitator [24] and significantly surpasses DiffPoseTalk [22]. Notably, unlike our approach and DiffPoseTalk [22], Imitator [24] employs a deterministic regression method that lacks the ability to produce diverse samples or offer editing capabilities. Additionally, we performed a motion editing experiment using DiffPoseTalk [22] which is shown in the suppl. video.

Method	Holistic synthesis	
	Face Motion (%)	Head motion (%)
1 Ours vs SadTalker [27]	88.13	86.43
2 Ours vs TalkShow [26]	90.77	87.96
3 Ours vs DiffPoseTalk* [22]	85.33	90.66

Table 1. User studies evaluating the naturalness of the facial and head motion. A total of 25 individuals participated in each A/B user-study. * : represents concurrent work.

Method	Style-similarity (%)
1 Ours vs Imitator [24]	55.64
2 Ours vs DiffPoseTalk* [22]	89.33

Table 2. User studies evaluating the style-similarity. A total of 25 individuals participated in each A/B user-study. * : represents concurrent work.

Impact of Guidance-Scale on facial motion synthesis: We investigate the impact of the classifier-free-guidance scale s [10] using the 'Lip-sync' and Div^L metrics on the non-personalized facial motion synthesis task. Lower guidance values yield animations with significantly more diverse motion but inferior lip-sync quality. Conversely, higher guidance values result in high-quality animation with reduced diversity. We observe a similar trend in our per-

ceptual evaluation. We find that the guidance scale s is an effective tool to increase synthesis diversity beyond all baselines with only a small loss of lip-sync accuracy for $0.3 \leq s \leq 1.0$.

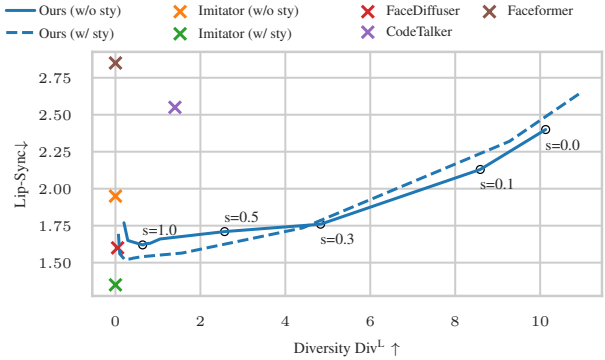


Figure 1. Evaluation of the impact of the classifier-free-guidance s on the facial motion synthesis task.

Impact of Noise: We conducted a noise sensitivity experiment similar to [2, 22, 24], where we added white noise to the input audio with a negative gain of 36db (low), 24db (medium), and 12db (high). As reported in Table 3, our method is robust to low(36db) and medium(24db) noise levels, produces facial motion with comparable quality. Please refer to the suppl. video for the qualitative results.

Method	$Div^L \uparrow$	Lip-Sync \downarrow
1 Ours (high noise)	6.41	2.56
2 Ours (med. noise)	2.54	1.97
3 Ours (low noise)	1.85	1.78

Table 3. Robustness to noise study with low, medium and high noise levels on the VOCaset [2].

Why standard diffusion-based head motion-editing fails? First, we analyze the standard diffusion based head motion editing and show why it fails. Then, we demonstrate how using our proposed Sparsely-Guided diffusion effectively addresses the issue and enables smoother head-motion editing. An illustration of a conditional inbetweening of a sequence across various diffusion steps t is shown in the Figure 2. Looking at the left column of the figure, it is clear that the standard diffusion mainly focuses on generating a valid sample from the distribution based on the

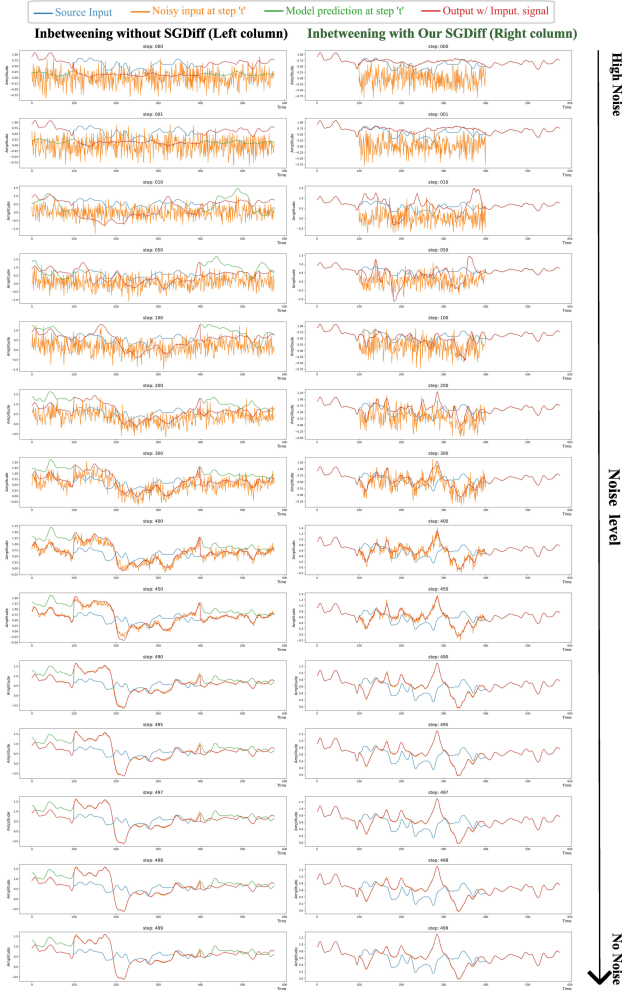


Figure 2. Illustration of a conditional head motion inbetweening of a sequence across various diffusion steps t with and without Sparsely-guided diffusion.

audio condition. As a result, it starts to ignore the imputation signal in the low noise regime, focusing instead on refining the sequence to produce an improved sample from the distribution. This approach results in jittery transitions when the imputation signal is replaced to generate the final inbetweened sample at the end of the sampling process. Throughout its training, the diffusion model was only trained to generate a valid sample from the distribution based on the audio condition, not to align with the imputation signal. Observing this, we introduced a sparsely-guided diffusion to incorporate guidance signals during training. This adjustment ensures that the diffusion model aligns with the imputation signal while still producing a sample from the distribution.

Is it possible to unconditionally synthesize and edit mo-

tion? While unconditional motion synthesis has been ex-

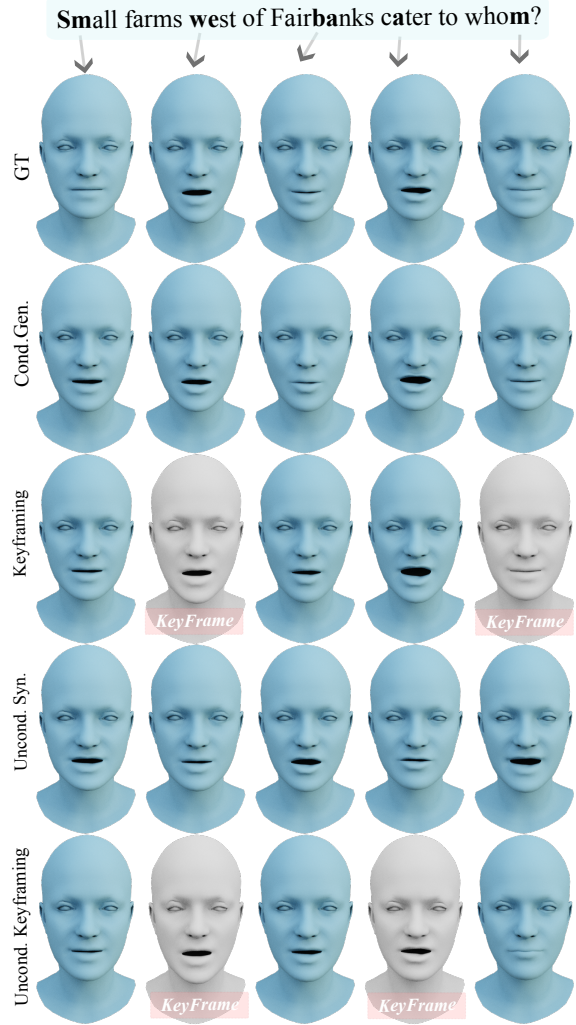


Figure 3. Qualitative illustration of facial motion inbetweening using our conditional (row 3) and unconditional model (row 5).

tensively applied in the motion synthesis domain [16, 23], to the best of our knowledge, its application in 3D facial animation synthesis remains widely unexplored. The significance of an unconstrained facial motion synthesis method cannot be overstated. It holds substantial potential for various applications, such as animating background characters in movies and games. Additionally, it enables targeted editing of specific facial elements—such as eye blinks and eyebrow motions—since these non-verbal facial expressions often exhibit weak or no correlation with audio features. Moreover, an unconditional model serves as a valuable motion prior for various downstream tasks, extending its utility beyond synthesis and editing applications. Our demonstration of unconditional synthesis and editing are showcased in Figure 3, underscoring the potential and versatility of such unconstrained models for 3D facial animation

synthesis. In Figure 3 (rows 2 and 3), we showcase a sequence synthesized conditionally and subsequently refined using keyframes. In Figure 3 (row 4), we present our unconditional synthesis results. As observed from the results, our model can unconditionally synthesize plausible facial motion. Additionally, in Figure 3 (row 5), we see that our method can unconditionally inbetween facial animation while preserving the speaking style of the target actor. This progression illustrates our model’s capabilities: from conditional synthesis and keyframe-based editing to unconditional synthesis and editing, while preserving the target actor’s speaking style. Please refer to the supplemental video for the study’s results in motion.

Impact of sparse guidance on Facial 3D animation synthesis: In this section, we evaluate the impact of using sparsely-guided diffusion (SGDiff) on 3D facial animation synthesis. To this end, we first trained the face motion generator with SGDiff, then personalized it with and without SGDiff using the VOCAsSet [2] test subject ‘0024’ and report the metrics in Table 4. While the introduction of sparse guidance improves diversity, it also reduces the lip-sync accuracy. The combination of window-based training and replacement of the input signal within the window acts as data augmentation, increasing the diversity and complexity of the training distribution. As a result, this leads to under-fitting, producing more generalized motion with high-diversity, but poorer lip-sync. As demonstrated in the main paper, style personalization is critical for 3D facial motion synthesis. Therefore, in our experiments, we employ a personalized facial motion model without sparse-guidance.

Method	Div ^L ↑	Lip-Sync ↓
1 Ours (without SGDiff)	1.35	1.4
2 Ours (with SGDiff)	2.19	2.44

Table 4. Impact of Sparsely-Guided diffusion on the 3D Facial motion synthesis. Introducing guidance during diffusion improves the diversity and reduces the lip-sync.

Impact of Personalization on Head-motion synthesis: We evaluate the impact of personalization on head-motion synthesis by personalizing the head-motion using the subjects in the HDTF test-set. The metric of this evaluation is reported in Table 6. From this experiment, we observe that personalization of the head motion often overfits to the speech context and cadence, resulting in high beat alignment and low diversity metrics. Since, one of our primary goal of this work is to enable diverse motion synthesis, we opted for a non-personalized head-motion model. This approach offers better diversity and more flexible editing capabilities.

Method	BA ↑	Div ^H ↑
1 Ours w/o. Personalization	0.338	0.007
2 Ours w. Personalization	0.673	0.002

Table 5. Impact of personalization on the head-motion synthesis.

Qualitative comparison against 3D facial motion synthesis methods: As mentioned in the main paper, we evaluate our method against the state-of-the-art methods VOCA [2], Faceformer [5], CodeTalker [25], EMOTE [4], FaceDiffuser [21] and Imitator [24] on facial motion synthesis task. A qualitative comparison to the facial motion synthesis baselines on a test sequence from the VOCAsSet is shown in Figure 4, where our method produces expressive facial animations that match the speaking style of the target subjects.

Quantitative evaluation on the BIWI Dataset: The main focus of our work is to enable diverse synthesis with precise control. This made both BIWI [6] and BEAT [14] incompatible for our study, as they are in different model spaces compared to the existing face trackers like [3, 7, 30]. This is a key necessity for personalization and subsequently face motion editing. More details about this is discussed in Section 2.2. Nevertheless, for completeness, we conduct a quantitative evaluation our 3D facial motion synthesis against the state-of-the-art methods trained on BIWI [6], in addition to the quantitative comparison presented on VOCA [2] in the main paper. To this end, we adopt the dataset setup used by [5, 21, 25] and only use the emotional sequence subset. Specifically, the data is split into a training set (BIWI-Train) containing 192 sentences and a validation set (BIWI-Val) with 24 sentences from 6 training subjects. There are two test sets: BIWI-Test-A, containing 24 sentences from seen subjects, and BIWI-Test-B, containing 32 sentences from 8 unseen subjects. For this experiment, we perform the qualitative study on the BIWI-Test-A and report the metric in Table 6. We use the Lip vertex error (*LVE*) metric used by the baselines [21, 25] in this experiment. From the results, we can observe that our method outperforms the baselines in terms of producing high-quality lip motion.

Method	LVE ↓
1 VOCA [2]	6.55
2 MeshTalk [18]	5.91
3 FaceFormer [5]	5.3
4 CodeTalker [25]	4.79
5 FaceDiffuser [21]	4.29
6 Ours	3.61

Table 6. Impact of personalization on the head-motion synthesis.

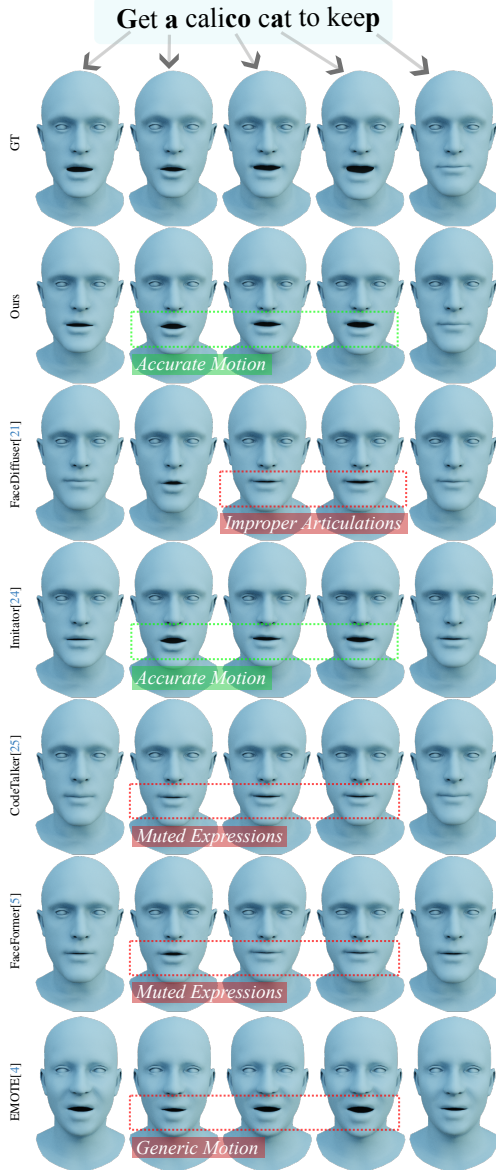


Figure 4. Qualitative comparison of the facial motion synthesis model.

2. Implementation

In this section, we provide more details on the diffusion model, dataset, baselines, and metrics.

2.1. Preliminaries

Denoising Diffusion Probabilistic Models: Our method is based on the diffusion framework of Sohl et al. [20], where a training sample x_0 gradually transforms into white noise through the addition of Gaussian noise across T steps.

This transformation is mathematically represented as:

$$x_t \sim q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), t = 1 \dots T, \quad (1)$$

where β_t is following a predefined variance schedule.

Following recent work [22, 23], we train a denoising model θ that can reverse this noisy diffusion and estimate the original sample x_0 from a noisy version x_t , guided by: $\hat{x}_0 = \theta(x_t, t, C)$. With θ being the neural network and C representing additional conditions. The reverse diffusion is achieved through:

$$q(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\theta(x_t, t, C), (1 - \bar{\alpha}_{t-1})I),$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{k=1}^t \alpha_k$.

To generate new samples, we start from random noise x_T and apply iterative denoising until reaching $t = 0$. We introduce diversity in generation using Classifier-Free Guidance (CFG) [10] by combining conditional and unconditional predictions of the network, controlled by a guidance scale s :

$$\theta_s(x_t, t, C) := \theta(x_t, t, \emptyset) + s \cdot [\theta(x_t, t, C) - \theta(x_t, t, \emptyset)],$$

adjusting s to balance between diversity and adherence to conditions. Following [11], the inverse diffusion process is then given through:

$$q(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\theta_s(x_t, t, C), (1 - \bar{\alpha}_{t-1})I), \quad (2)$$

with $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{k=1}^t \alpha_k$. For generating new samples, we randomly sample x_T from a Gaussian distribution and iteratively denoise it until $t = 0$ is reached.

To add diversity, we employ Classifier-Free Guidance (CFG) [10] and calculate the output as a weighted sum of the conditional and unconditional prediction:

$$\theta_s(x_t, t, C) := \theta(x_t, t, \emptyset) + s \cdot [\theta(x_t, t, C) - \theta(x_t, t, \emptyset)], \quad (3)$$

where s is the guidance scale and $\theta(x_t, t, \emptyset)$ denotes the unconditional prediction in which we set the audio conditions to zero. Note that while CFG is typically used with a guidance scale > 1 to enhance alignment with the condition, we set it to values < 1 (0.5 unless specified otherwise) to increase diversity.

2.2. Dataset

VOCA: We train our facial motion model on the VOCAset [2] since it provides high-quality, speech-aligned 3D face scan sequences. It consists of 12 actors (6 female and 6 male) with 40 sequences each with a length of 3-5 seconds, resampled at 30fps. Following previous work [24], we use the train/val/test set split of 8, 2, 2 actors. All 40 sequences of the training actors are used during training. However, for the test and validation, only 20 sequences that do not

overlap with the speech scripts of the training sequences are used. For the style adaption experiment, we split the 40 sequences of the test actors to 18, 2, 20 for train/val/test sets. The test sequences of the experiments w/ and w/o style adaptation are identical, allowing a direct comparison of the scores in the quantitative comparison in the main paper (Table 2).

In-the-wild dataset: We evaluate person-specific fine-tuning on in-the-wild video sequences from Imitator [24]. The provided videos are 2 minutes long which we divide into 60/30/30 seconds for train/val/test respectively. Similar to Imitator, we employ the MICA tracker [30] to extract the face motion tracking for the personalization step.

HDTF: We train our head-motion generator on the HDTF [28] dataset. The High-definition Talking Face Dataset (HDTF) is a large in-the-wild audio-visual dataset for talking face generation. It consists of about 362 different high-resolution (720P or 1080P) YouTube videos of 15.8 hours in total. Using the download and processing script provided by the authors, we extracted 352 videos with 246 unique subjects. We additionally crop the video to 30 seconds long and use them for extracting head-poses using the MICA tracker [30], which provides head poses as global axis rotation. For our experiments, we split the dataset into 300/20/32 sequences for train/val/test accordingly.

Discussion: In this work, we employ the VOCAs, HDTF, and Imitator’s in-the-wild dataset to train our method for generating and editing 3D facial animations with head motion. The motivation of generating and editing holistic 3D facial animation made both BIWI [6] and BEAT [14] incompatible for our study, both BIWI and BEAT are in different model spaces compared to the existing face trackers like [3, 7, 30], which is a key necessity for personalization and subsequently face motion editing. Such a problem could in theory be addressed by converting the meshes provided in the dataset to our target FLAME model space by optimization-based fitting using pre-defined correspondence between the source and target mesh space. However, for BIWI the noisy surface reconstructions provided in the dataset and incomplete face models make the fitting challenging and reduces the quality of the fitted meshes further. Similarly, for BEAT, the dependence on ARKit which produces improper lip-closures and not fully completed face model, reduces the realism of the reconstructed sequences. As studied in [24], lip-closures are paramount in conveying realism for the generated sequences.

2.3. Baselines

Holistic 3D motion synthesis: For TalkShow [26], we use the pre-trained model provided in their official repos-

itory and extract the predicted facial and head motion parameters for our evaluation. For SadTalker [27], we use the pre-trained model provided in the repository to generate 2D talking face videos and use the MICA tracker [30] to the face and head motion. For DiffPoseTalk [22], we use the pre-trained provided in their repository. Similar to our work, DiffPoseTalk [22] requires 3D reconstructions of target subjects for style-personalization. For 3D reconstruction, DiffPoseTalk utilize their own customer-tracker, which, at the time of the submission was not available in their repository. Hence, we substitute their tracker with EMOCaV2 [3], a publicly available face tracker in our experiments.

Facial Motion Synthesis: For VOCA [2], Faceformer [5], Imitator [24] and FaceDiffuser [21], we use the pre-trained model provided in the official repositories. For CodeTalker [25], we adapt the official implementation to add the functionality of generating diverse motion. Especially, we re-train the audio-conditioned codebook sampling (stage 02) to randomly sample a code from the top ‘m’ closest codes instead of always using the closest code. This process is in spirit close to training the language-based models, where a new diverse text sequence is generated by sampling the 2nd or 3rd closest language token over the token with maximum probability. By adapting this method, we ensure that CodeTalker can generate diverse samples for a given audio input. For EMOTE [4], we request the authors to run their method on the VOCAs [2] and use it for the qualitative and perceptual user study.

2.4. Training Details

Facial Motion Synthesis: We train our method using ADAM [13] with a learning rate of $1e-4$ for 140K iterations with a batch size of 64. Our diffusion framework is based on the Gaussian diffusion from Nichol *et al.* [15], we set the diffusion step to 500 for our experiments. During training, we randomly crop the sequences to the length of 30 frames. Our lightweight architecture enables us to train our model on a single Nvidia Quadro P6000 32GB within 30 hours. The lightweight architecture is also critical for person-specific style adaptation with a short reference video. For person-specific speaking style, we use the same training setup as from the generalized setting, except that we only train it for 30K iterations. For evaluating the best checkpoint, we fix the guidance scale $s = 0.99$ and evaluate all the saved checkpoints on the validation set. Further, we fix the best checkpoint and vary the guidance scale from $s=0, 0.1 \dots$ to 1.0 with an increment of 0.1 and find the best guidance factor. From our experiment, we found the guidance scale of 0.5 balances the lip-synchronization and diversity and provides the best results.

Head Motion Synthesis Similar to the Facial motion synthesis pipeline, we train our method using ADAM [13] with a learning rate of $1e-4$ for 100K iterations with a batch size of 64. During training the sequences are randomly cropped to 300 frames long. For our inbetweening and keyframing-based Guided motion model training, we randomly sample a mask of arbitrary length for imputation signal, using which the noisy input is replaced with the ground truth imputation signal.

2.5. Inference

Our method takes 3.15 sec to produce 1 sec (30 frame) of facial motion and 1.04 sec to produce 1 sec (30 frame) of head motion on a single Nvidia GeForce RTX 3090 24GB, compared to 5.78 sec for the concurrent method FaceDiffuser [21]. In total, our method takes 4.19 sec to produce 1 sec (30 frame) of holistic 3D facial animation, compared to 6.78 sec for TalkSHOW [26].

2.6. Metrics

Lip-Sync measures the lip synchronization using Dynamic Time Warping to compute the temporal similarity [24].

Diversity metric introduced by Ren et al. [17] measures the diversity of 3D motions for the same text input. We employ this metric and propose Div^L and Div^H to measure the diversity of lip motion and head motions generated from the same audio. Given a set of generated 3D facial or head motions with N sequences generated from the same audio condition. The diversity can be formalized as:

$$Diversity = \frac{1}{L} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \|m_i - m_j\|_2 \quad (4)$$

Where m_i represents the i -th motion and L is the total number of possible combinations in the generated motion set.

Beat alignment (BA) : Similar to DiffPoseTalk [22], we employ a modified beat alignment BA to measure the synchronization of the head movement beats between the predicted and ground truth motion, where we calculate the average temporal distance between beat in predicted head movement its closest ground truth beat as the Beat Align Score.

$$Beat\ Align\ Score = \frac{1}{|\mathbf{B}_g|} \sum_{t_g \in \mathbf{B}_g} \exp \left(-\frac{\min_{t_p \in \mathbf{B}_p} \|t^p - t^g\|_2^2}{2\sigma^2} \right) \quad (5)$$

Where \mathbf{B}_g and \mathbf{B}_p record the time of the beats in the ground truth and predicted head motion respectively, while σ is the normalized parameter which is set to be 3 in our experiment.

Discussion The $L2$ -based vertex error metrics employed in previous studies [5, 24, 25] are not apt for our task due to its preference for solutions that are close to the mean of the dataset, which penalizes the diversity present in our predictions.

2.7. Perceptual Study

We conducted A/B user studies to assess our method’s perceptual performance. First, we conducted a study to evaluate the holistic motion synthesis based on the naturalness of the facial and head motion to the input audio. For this, we sampled 10 sequences from the test set of the HDTF and 10 external audio from YouTube and synthesized holistic 3D facial motion using our method and the baselines [22, 26, 27] resulting in a total of 60 A/B comparisons including ground truth. For extracting the ground truth for the YouTube sequences, similar to the HDTF dataset processing we utilize the MICA tracker [30] to extract the facial and head motion. Through Amazon Mechanical Turk(AMT) and Google forms, we divided the A/B comparisons into 3 HITs (Human Intelligence Task), each with 25 individual assignments. For each HIT, users were instructed to select their preference for a method based naturalness of the head and facial motion with synchronization. Second for facial motion synthesis, we sample 20 sequences combined from the VOCaset test set and the in-the-wild sequences from Imitator, resulting in 100 A/B comparisons across five baselines. On Amazon Mechanical Turk(AMT), we divided the A/B comparisons into 5 HITs (Human Intelligence Task), each with 25 individual assignments. For each HIT, users select their preference for a method based on expressiveness and lip-synchronization. Finally, we evaluated the speaking style preservation of our personalized model in comparison to Imitator. To this end, the AMT users rated the similarity based on a reference video and the synthesized videos of the VOCA test set. Figure 5 illustrates an example interface in our user-study.

3. Ethical Impact

We introduce a method for realistic facial animation synthesis and editing that matches the speaking style of any given target actor. These animations hold promise for driving virtual avatars in AR or VR settings, especially, in immersive communication technologies. Yet, it is essential to acknowledge the potential pitfalls of such advancements, notably in the realm of ‘DeepFakes.’ By employing voice cloning techniques, our method can generate 3D facial animations that drive digital avatar methods like [1, 8, 9, 12, 29], which could be abused for identity theft, cyberbullying, and various criminal activities. Advocating for transparent research practices, we strive to illuminate the risks associated with technology misuse. Sharing our implementation aims to foster research in digital multimedia forensics, particularly in developing synthesis methods

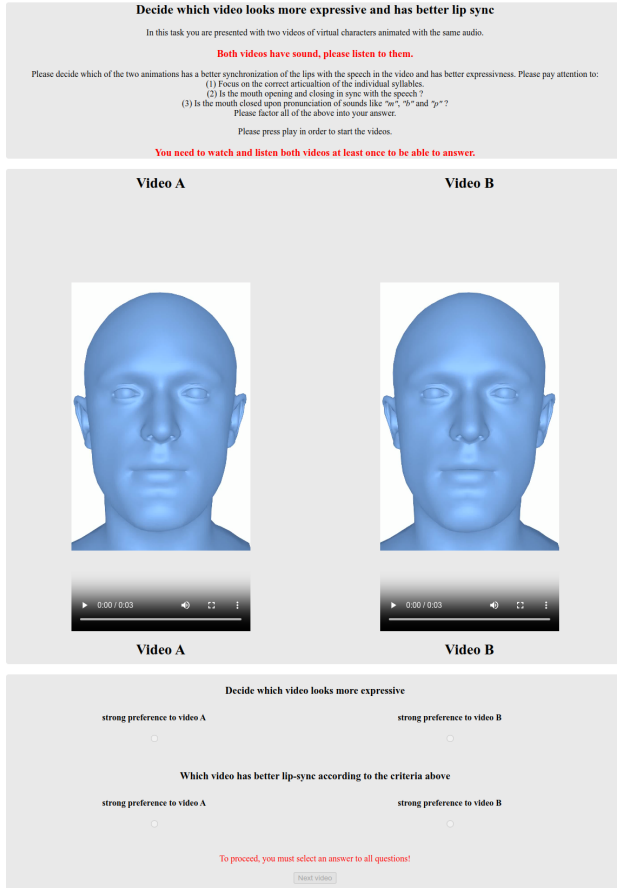


Figure 5. Example of the interface employed for our user-study.

crucial for training data utilized in spotting forgeries [19].

References

- [1] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. FLARE: Fast learning of animatable and relightable mesh avatars. *ACM Transactions on Graphics*, 42:15, 2023. 6
- [2] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, Learning, and Synthesis of 3D Speaking Styles. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10093–10103, Long Beach, CA, USA, 2019. IEEE. 1, 3, 4, 5
- [3] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 3, 5
- [4] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. *ACM*, 2023. 3, 4, 5
- [5] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 3, 4, 5, 6
- [6] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591 – 598, 2010. 3, 5
- [7] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos, 2022. 3, 5
- [8] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 6
- [9] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 6
- [10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 1, 4
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4
- [12] Berna Kabadayi, Wojciech Zielonka, Bharat Lal Bhatnagar, Gerard Pons-Moll, and Justus Thies. Gan-avatar: Controllable personalized gan-based human head avatar. In *International Conference on 3D Vision (3DV)*, 2024. 6
- [13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5, 6
- [14] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297*, 2022. 3, 5
- [15] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 5
- [16] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data, 2022. 2
- [17] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model, 2023. 6
- [18] Alexander Richard, Michael Zollhofer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1153–1162, Montreal, QC, Canada, 2021. IEEE. 3
- [19] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. *ICCV 2019*, 2019. 7
- [20] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using

nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

4

- [21] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France*, New York, NY, USA, 2023. ACM. 3, 4, 5, 6
- [22] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Gaetan Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong jin Liu. Diff-posetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models, 2023. 1, 4, 5, 6
- [23] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4
- [24] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20621–20631, 2023. 1, 3, 4, 5, 6
- [25] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 3, 4, 5, 6
- [26] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 1, 5, 6
- [27] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. *arXiv preprint arXiv:2211.12194*, 2022. 1, 5, 6
- [28] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 5
- [29] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. 6
- [30] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. *ECCV*, 2022. 3, 5, 6