THMM-DiT: Talking Head Motion Modeling with Diffusion Transformer

Sejong Yang*
Yonsei University
sejong.yang@yonsei.ac.kr

Seoung Wug Oh
Adobe Research
seoh@adobe.com

Yang Zhou
Adobe Research
yazhou@adobe.com

Seon Joo Kim
Yonsei University
seonjookim@yonsei.ac.kr

Abstract

Generating natural talking head motion is a challenging task due to the one-to-many nature of speech-to-motion mapping, the high dimensionality of RGB video, and the difficulty of modeling dynamic head poses. In this technical report, we propose a new approach to generating natural talking head motion that addresses these challenges. Our approach uses a diffusion model to generate a distribution of possible head poses, which is then conditioned on the given audio to produce a natural-looking talking head. We also use a face template to reduce the computational resources required to generate high-quality RGB videos. Finally, we employ long clue frames with spatio-temporal attention of transformer to generate natural long-term sequences of head poses. Our approach is able to generate dynamic head poses in the long term while accurately synchronizing mouth shapes with the given audio.

1 Introduction

Motion modeling is a challenging task, but Human Motion Diffusion Model (MDM) [16] has shown good performance with SMPL [10] and Diffusion Models [12]. Unlike body motion modeling, there is no annotated dataset for face motion. Therefore, we utilize 3DMM [13] reconstruction network 3DDFA_v2 [5] with talking face video dataset VoxCeleb2 [2] and Celebv-Text [19]. This works well for unsupervised talking motion generation, but there are two problems for conditional generation with speech.

First, in-the-wild talking face video has biased speech distribution unlike GRID audio-visual speech corpus. To solve this problem, we utilize WHISPER to extract semantic speech features and concatenate them with melspectrograms from the LombardGrid dataset. For in-the-wild and high-quality generation, we also preprocess VFHQ and merge them as a combined dataset.

Second, the MotionDiffuse [20] architecture only has temporal attention because text and motion modalities have different lengths in general. To utilize spatial attention, we change our face template to facial landmarks and revise the attention block. We evaluate our method on VoxCeleb2 [2], LombardGrid [1], VFHQ [17], and in-the-wild data samples.

*This work is done as collaboration of Yonsei and Adobe

Our results show that our method can generate high-quality talking motion with speech audio. We also show that our method can generate talking motion from in-the-wild data, which is a challenging task due to the bias in the speech distribution.

2 Related Works

2.1 Diffusion Models

Recent diffusion models have shown great performance and a wide range of applications. [6, 12, 7, 15, 14] Diffusion models learn the exact data distribution with exact likelihood [12], which is known to achieve better precision and recall than VAE [8] or GAN [4]. Unlike the Flow-based Model [9], diffusion models are not limited by architecture, so they can be widely used in various fields. In this study, we aim to address the one-to-many nature of speech-conditioned talking head motion generation by exploiting the ability of diffusion model to achieve fidelity and diversity simultaneously.

2.2 Talking Motion Modeling with Template Model

Previous work has explored the use of face templates for speech-to-motion generation. For example, FaceFormer [3] generated 3DMM-based talking motion in an auto-regressive manner using transformers. UniFLG [11] generated talking motion in Japanese based on facial landmarks. However, both methods were trained on limited datasets and had difficulty responding to in-the-wild speech.

In addition to speech-to-motion generation, researchers have also explored the use of face templates as an intermediate representation for speech-driven talking video generation. GeneFace [18] and IP_LAP [22] used facial landmarks, while SadTalker [21] used 3DMM. However, these studies were limited by the use of mostly static datasets, which hindered the generation of dynamic head poses.

3 Method

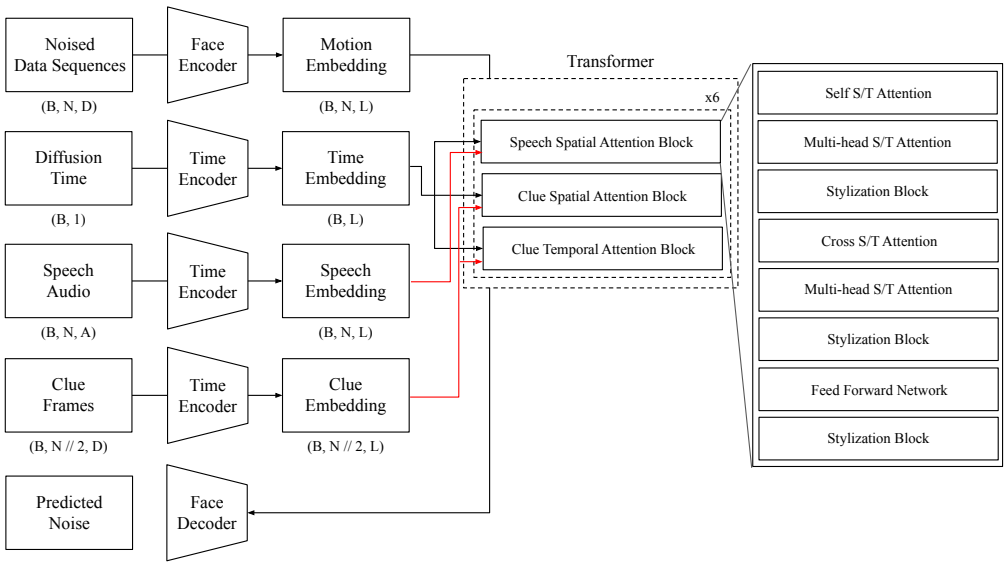

Figure 1: The Overview of Model Architecture.

Our model builds upon ImprovedDDPM [12] following the MotionDiffuse [20]. The overview of the model architecture is shown in Figure 2. Each transformer block can use either spatial or temporal attention, as indicated by the S/T labels in the right detail box. The stylization block receives an additional embedding that is the sum of the time embedding and the temporally averaged condition embedding, such as the speech of the clue. The cross attention block uses each condition embedding

individually, except for the clue spatial attention block, which uses the temporally last clue embedding for spatial attention.

The method we propose in this study differs from previous works in two key aspects. Firstly, our approach incorporates a diffusion transformer to effectively handle the one-to-many nature of speech-to-motion generation. By leveraging diffusion models, we are able to generate high-fidelity samples with enhanced diversity through the utilization of clue frames for next prediction, rather than relying on reference frames for lip region editing. Secondly, our model introduces the utilization of long clue frames instead of a single clue frame to extract motion hints. While previous works solely extracted appearance hints from a single clue frame, our model utilize the power of motion hints, enabling the generation of dynamic talking head motion.

4 Experiment

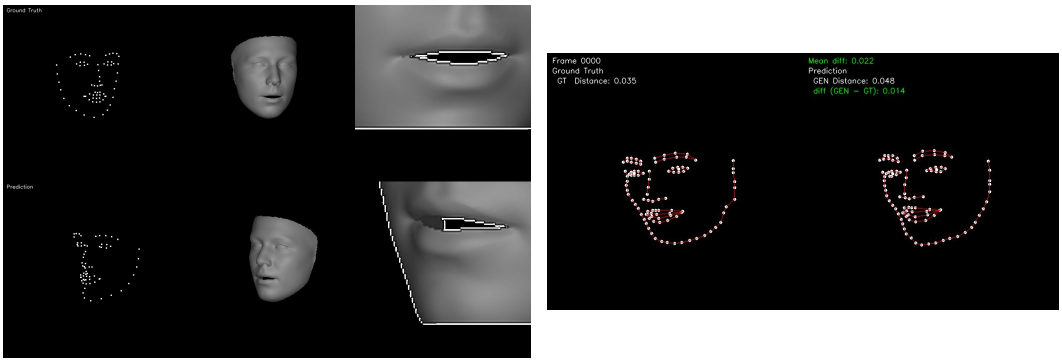

Figure 2: The GT and generated result of our model. The left is unsupervised 3DMM sequence generation and the right is speech conditional landmark sequence generation.

Unsupervised 3DMM Sequence Generation The model is trained and tested on VoxCeleb2 dataset. In this setup, we utilize the self temporal attention block because there is no condition. More results can be found on `https://drive.google.com/drive/folders/1V24yKKAB8f4Wls5vLZ74-qUTG2ysNCMX`.

Speech Conditional Landmark Sequence Generation The model is trained and tested on merged dataset of Lombardgrid and VFHQ dataset. We calculate the distance between upper lip and lower lip with landmark that we call lip distance. Then the difference of lip distance between GT and Generated is measured to check the alignment of speech and lip motion. The difference of lip distance was 0.014 for Lombardgrid and 0.020 for VFHQ. More results for lombardgrid dataset can be found on `https://drive.google.com/drive/folders/1BcvCljzFfXzPA6iY7Lfs_rMo2uxZezEJ`. More results for VFHQ dataset can be found on `https://drive.google.com/drive/folders/1iVgXxgAAM_GHyTBo5ZF1GsVszsgfAInG`.

5 Conclusion

This technical report introduces our proposal for talking head motion modeling using the diffusion transformer. We emphasize the importance of employing spatial-temporal attention in an effective manner, specifically by leveraging long clue frames, as a crucial aspect for achieving successful diffusion transformer training. We firmly believe that our findings can be extended to facilitate the generation of even more naturalistic talking head motion.

References

[1] Alghamdi, N., Maddock, S., Marxer, R., Barker, J., Brown, G.J.: A corpus of audio-visual lombard speech with frontal and profile views. The Journal of the Acoustical Society of America (2018)

[2] Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. International Speech Communication Association (INTERSPEECH) (2018)

[3] Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM (2020)

[5] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Proceedings of Proceedings of European Conference on Computer Vision (ECCV) (2020)

[6] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) (2020)

[7] Ho, J., Salimans, T.: Classifier-free diffusion guidance. Advances in Neural Information Processing Systems (NeurIPS) (2021)

[8] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. International Conference for Learning Representations (ICLR) (2014)

[9] Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. Advances in Neural Information Processing Systems (NeurIPS) (2018)

[10] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) (2015)

[11] Mitsui, K., Hono, Y., Sawada, K.: Uniflg: Unified facial landmark generator from text or speech. International Speech Communication Association (INTERSPEECH) (2023)

[12] Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning (ICML) (2021)

[13] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance (2009)

[14] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[15] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. International Conference for Learning Representations (ICLR) (2021)

[16] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. International Conference for Learning Representations (ICLR) (2023)

[17] Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[18] Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. International Conference for Learning Representations (ICLR) (2023)

[19] Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: CelebV-Text: A large-scale facial text-video dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

[20] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)

[21] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: 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 (CVPR) (2023)

[22] Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)