EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions (*Supplementary Material*)

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo

Institute for Intelligent Computing, Alibaba Group
{tianlinrui.tlr, wilson.wq, zhangbang.zb, liefeng.bo}@alibaba-inc.com
https://humanaigc.github.io/emote-portrait-alive/

This supplementary material contains additional information that could not be included in the main manuscript due to space limitations. It includes a wider exploration of our method's results, and we discuss about our training dataset. Following that, we provided a candid discussion on the limitations of EMO, and a preview of future research directions.

A More results

A.1 User study

Table 1: User study.

	lip sync	vividness
Wav2Lip [6]	3.90	1.66
SadTalker [10]	3.38	2.34
DreamTalk [4]	4.05	3.22
MakeItTalk [12]	1.66	2.66
Ours	**4.17**	**4.38**

We also carried out a user study using the generated outputs, involving 20 participants, evenly split between 10 males and 10 females, ranging in age from 20 to 60, and with diverse levels of computer technical expertise. For every volunteer, We will show them the results of all methods on the same image and audio simultaneously in each round. The volunteers are asked to rate each video between 1 and 5 (higher is better), in terms of lip synchronization and vividness. The results are shown in the Table 1 and our method significantly outperforms other methods, especially on the vividness.

10
20
35
40
50

Fig. 1: The generated results under different denoising steps.

A.2 Motion frames

Our method employs motion frames to bolster consistency across clips generated in sequence. Some other video generation techniques [2] may employ 'frame replacing'-substituting the initial m frames with the final m frames from the preceding clip at each denoising step and using temporal modules for frame-to-frame coherence. Our approach does not utilize strong control signals, such as pose sequence, to guide the sequential generation process of video clips. Thus, along with a sensitivity to 'jump cuts' highlighted in Sec B.2, our method might experience frame discontinuity during clip transitions.

A.3 Inference steps

In our experiments, we observed that the number of denoising steps during inference critically influences the quality of the generated output. As depicted in the Figure 1, a suboptimal number of steps (fewer than 20) results in temporal inconsistencies across frames, as well as a prevalence of visual artifacts within the sequences. When the number of denoising steps is marginally increased (within the range of 20–35), the artifacts are somewhat ameliorated, however, the generated characters might exhibit noticeable jitter and instability. Our experiments indicate that employing more than 35 denoising steps enhances stability and temporal coherence, leading to more reliable results in the synthesized sequences.

B Dataset

B.1 Data overview

Our training dataset mainly contains 1) HDTF [11], encompassing 15.8 hours of high-fidelity talking head videos, featuring approximately 362 unique character identities with a majority being anchors and spokesmen; 2) VFHQ [9], which contains 16k high-resolution talking video clips without audio; 3) A diverse collection of speech and singing data aggregating to 250 hours, sourced from online platforms. The videos are publicly accessible, and the collectors are also instructed to carefully review the content to exclude any personally identifiable information, thus ensuring adherence to privacy standards and the terms of use of the platforms. Our dataset features over 10,000 unique character identities, with a focus on English and Chinese languages. We ensured that the character's position, camera angle, and background in the collected data remained relatively unchanged, with each frame capturing a single individual. Unlike some talking head methods that rely on Voxceleb [4, 5, 8], we opted not to use it due to its frequent centering on facial centroids, leading to unstable camera movements.

However, as indicated in Table 1 of the main manuscript, our method still exhibits exceptional performance in the absence of our self-collected dataset. Our model yields satisfactory results when trained solely on publicly available datasets. The extensive dataset we compiled primarily enhances facial expressions and video content dynamics.

B.2 Preprocessing and labeling

Preprocessing the orginal videos. Discontinuities in training data, such as inconsistencies in character appearance and camera switches, pose significant challenges for model training, often resulting in the generation of unstable videos. To mitigate these effects, our approach involves segmenting videos into shorter clips that maintain both temporal coherence and scene consistency, similar to the method described in SVD [1]. By employing PySceneDetect for scene transition identification, we ensure that each clip, ranging from 3 to 12 seconds, contributes to a more reliable and stable dataset for training our generative models. Furthermore, 'jump cuts', prevalent in speech videos for maintaining narrative continuity, present additional detection challenges due to their subtle nature. Our solution incorporates 'speed layers' to address the abrupt velocity changes associated with 'jump cuts', thereby enhancing the fluidity and consistency of the generated video content.

Labeling the data. We performed cropping on the video clips based on the expanded facial bounding boxes of the characters within the clips, and converted each clips to 30 FPS. And we label the cropped clips with 1) deploying MediaPipe [3] to ascertain the facial bounding box in all frames, thereby delineating the facial regions; 2) extracting audio embeddings using the pre-trained Wav2Vec model [7]; and 3) determining the character's 6-DoF (six degrees of freedom) head pose to calculate frame-by-frame velocities.

Fig. 2: Exhibition of Artifacts: This includes the manifestation of subtitle-like patterns and inaccurately rendered body parts. Notably, in certain instances, the generated frames may self-correct in subsequent sequences.

C Limitations and future work

As introduced in the main manuscript, our model does not employ explicit control signals to control the character's motion. As shown in the Figure 2, there is a propensity for the model to inadvertently generate body parts into frames, particularly when the audio input features prominently expressive emotions. This tendency arises from the nature of our training dataset, where characters exhibiting pronounced emotional states are frequently associated with more dynamic hand and body movements. Given that our primary focus lies on the head region, the dataset is deficient in data pertaining to other body parts, with a mere 3% of frames featuring hands. In instances where the model attempts to render these body parts unintentionally, the result is often the generation of incorrect body parts, leading to artifacts.

Additionally, the model may produce caption-like patterns under some circumstances. This issue stems from a subset of the training data sourced from the internet that includes videos with embedded subtitles, hence introducing unwanted textual artifacts into the generated frames. This phenomenon has been observed and is not exclusive to our model; it is also prevalent in the output of Text-to-Image (T2I) models.

To address these issues, one potential solution involves the introduction of control signals for body parts and subtitles. More specifically, utilizing mask-like input, similar to face regions.

Another limitation of our model, EMO, is its reliance on audio as the principal control signal. EMO has been trained to learn the correlation between tonal features in the audio and facial expressions in characters. However, this association can result in expressions for the driven characters that may not always align with user expectations, limiting the ability to produce desired video outcomes in a controlled manner. Introducing a mechanism to define emotions could enhance user convenience by providing more predictable results.

Compared to the other diffusion-free talking head generation methods, EMO is more time-consuming, it generates 12 frames (one clip) per 18 seconds (under 40 denoising steps) on A100 GPU.

Despite these challenges, EMO represents a significant leap forward in the development of highly expressive video generation models, laying a foundation for future innovation in this field. We leave these issues as open questions for subsequent research.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 3
2. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023) 2
3. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines (06 2019) 3
4. Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., Deng, Z.: Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. arXiv preprint arXiv:2312.09767 (2023) 1, 3
5. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: Large-scale speaker verification in the wild. Computer Science and Language (2019) 3
6. Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 484–492. MM '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394171.3413532, https://doi.org/10.1145/3394171.3413532 1
7. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. pp. 3465–3469 (09 2019). https://doi.org/10.21437/Interspeech.2019-1873 3
8. Sun, X., Zhang, L., Zhu, H., Zhang, P., Zhang, B., Ji, X., Zhou, K., Gao, D., Bo, L., Cao, X.: Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. arXiv preprint arXiv:2312.01841 (2023) 3
9. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022) 3
10. 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: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8652–8661. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.00836, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00836 1
11. Zhang, Z., Li, L., Ding, Y., Fan, C.: 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. pp. 3661–3670 (2021) 3

12. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makeittalk: Speaker-aware talking-head animation. ACM Transactions on Graphics **39**(6) (2020) 1