

Supplementary Materials for FLOAT: Generative Motion Latent Flow Matching for Audio-driven Talking Portrait

Taekyung Ki^{1*} Dongchan Min¹ Gyeongsu Chae²

¹KAIST ²DeepBrain AI Inc.

{taekyung.ki, alsehdcks95}@kaist.ac.kr gc@deepbrain.io

<https://deepbrainai-research.github.io/float/>

In this supplement, we first provide more details on motion latent auto-encoder in Appendix A, regarding the model itself (Appendix A.1), methods for improving the fidelity of facial components (Appendix A.2), the training objective (Appendix A.3), and implementation details (Appendix A.4).

In Appendix B, we provide more details on FLOAT, regarding details on evaluation metrics (Appendix B.1), baselines (Appendix B.2), and ablation studies (Appendix B.3).

In Appendix C, we provide additional results, including comparison results (Appendix C.1), out-of-distribution results (Appendix C.2), and user study (Appendix C.3).

Finally, we discuss ethical considerations, limitations, and future work in Appendix D.

A. More on Motion Latent Auto-encoder

In this section, we provide more details on our motion latent auto-encoder, including its model architecture, dataset, and training strategy.

A.1. Model

We provide a detailed model architecture of our motion latent auto-encoder in Fig. 8.

In Fig. 2a, Fig. 2b, Fig. 2c, and Fig. 2d, we present visualization results of the latent decomposition

$$w_S = w_{S \rightarrow r} + w_{r \rightarrow S} \in \mathbb{R}^d \quad (1)$$

of a source image S , following the approach of [35]. Notably, the identity latent $w_{r \rightarrow S}$ is decoded into image featuring the average head pose, expression, and field of view in pixel space.

A.2. Improving Fidelity of Facial Components

Facial Components: Texture vs. Structure As highlighted in face restoration work [34], facial components

*This work was done during South Korea Mandatory Military Service at DeepBrain AI Inc.

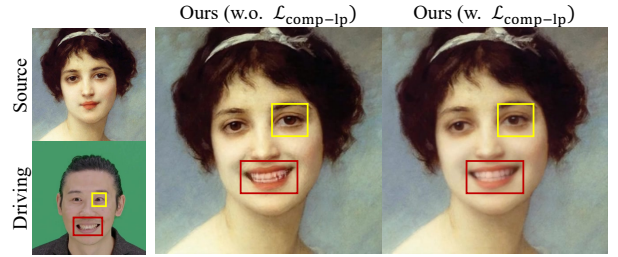


Figure 1. Ablation study on Facial Component Loss $\mathcal{L}_{\text{comp-lp}}$. It significantly improves the image fidelity of facial component (e.g., teeth, highlighted in red box) and fined-grained motion (eyeball movement, highlighted in yellow box).

such as eyeballs and teeth play a important role in the perceptual quality of generated images. It treats the issue as a lack of *texture* (lying in high frequencies) and mitigate it by introducing facial component discriminators with the gram matrix statistics matching. This approach is appropriate in face restoration, where training objective is to reconstruct a clear image from a degraded one that maintains the same spatial structure, ensuring that the low-frequency structure preserved.

However, in the context of training a motion auto-encoder, spatial mismatches are inevitably involved. Therefore, naively applying such discriminators proves ineffective. Instead, achieving high-fidelity facial components in a motion auto-encoder is more closely related to structural problems (lying in low frequencies) than to texture issues as shown in Fig. 2f.

Facial Component Perceptual Loss $\mathcal{L}_{\text{comp-lp}}$ We introduce a simple yet effective *facial component perceptual loss*, which leverages the standard perceptual loss \mathcal{L}_{lp} [41] known for its ability to capture structural features lying in low frequencies. Formally, the facial component perceptual

Table 1. Quantitative comparison result (Same-identity) of motion latent auto-encoders on HDTF [43] / RAVDESS [19] / VFHQ [36]. The best result for each metric is in **bold**.
[†]: Results generated by official implementation (256 × 256)

Method	FID ↓	FVD ↓	LPIPS ↓	E-FID ↓	P-FID ↓
LIA [†] [35]	47.481 / 67.541 / 89.209	172.195 / 130.836 / 342.964	0.184 / 0.122 / 0.245	1.279 / 1.153 / 1.106	0.120 / 0.005 / 0.013
Ours (w.o. $\mathcal{L}_{comp-lp}$)	21.061 / 28.866 / 46.950	150.340 / 103.145 / 299.757	0.110 / 0.072 / 0.165	1.369 / 1.157 / 0.872	0.011 / 0.010 / 0.014
Ours	19.803 / 23.350 / 43.992	147.089 / 100.345 / 291.560	0.108 / 0.062 / 0.161	1.334 / 1.053 / 1.006	0.010 / 0.008 / 0.012

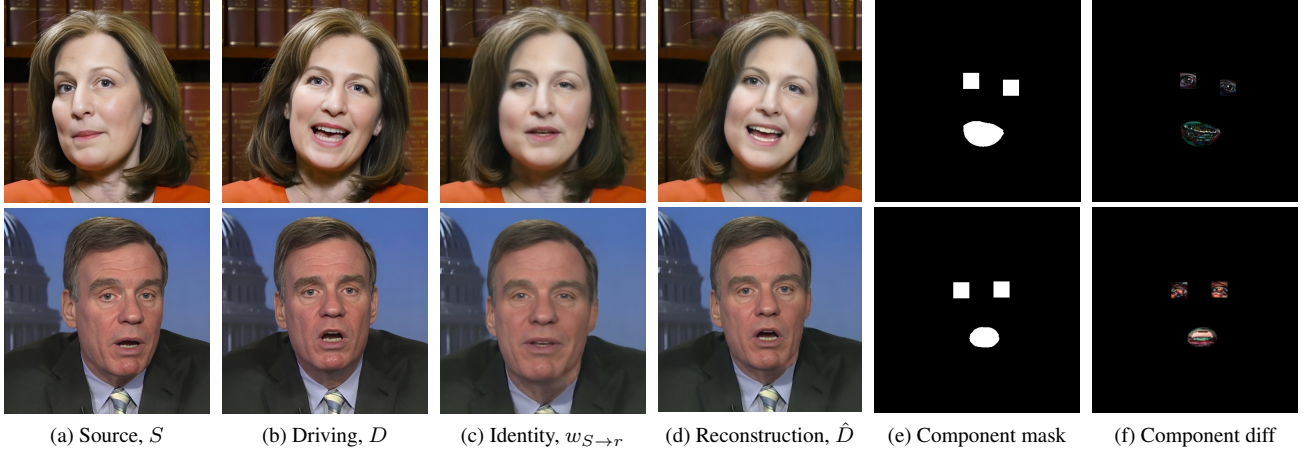


Figure 2. Visualization results of the motion latent auto-encoder.

loss is defined by

$$\sum_{i=1}^N \frac{1}{|M_i|} \|M_i \otimes \phi_i(\hat{D}) - M_i \otimes \phi_i(D)\|_1, \quad (2)$$

where D is the driving, \hat{D} is the generated image, N is the number of feature pyramid scales, $\phi_i(X)$ is the i -th feature of the input image X computed by VGG-19 [24, 41], M_i is the binary mask of the facial components that has same size with $\phi_i(X)$, and $|M_i|$ is the sum of all values in the binary mask M_i . We adopt a single perceptual loss with $N = 4$ scales of VGG-19 feature pyramids. It is worth noting that we mask all the multi-resolution features (not only the image).

To compute the facial component mask M_i , we utilize an off-the-shelf face segmentation model [39] for tight mouth regions and face landmark detector [1] for the bounding box regions of the eyes as illustrated in Fig. 2e.

In Tab. 1, we conduct ablation studies on motion latent auto-encoders. Notably, $\mathcal{L}_{comp-lp}$ is consistently improves the image fidelity over three datasets. As illustrated in Fig. 1, an additional advantage of $\mathcal{L}_{comp-lp}$ is its ability to directly supervise fine-grained motion (often neglected due to large head motion) such as eyeball movement without any external driving conditions such as eye-gazing direction [7].

A.3. Training Objective

We train our motion latent auto-encoder by reconstructing a driving image D from a source image S , both sampled from

the same video clip.

The total loss function \mathcal{L}_{total} for the motion latent auto-encoder is defined as

$$\begin{aligned} \mathcal{L}_{total} = & \mathcal{L}_{L1} + \lambda_{lp} \mathcal{L}_{lp} + \lambda_{comp-lp} \mathcal{L}_{comp-lp} \\ & + \lambda_{full-adv} \mathcal{L}_{full-adv} \\ & + \lambda_{eye-adv} \mathcal{L}_{eye-adv} + \lambda_{eye-FSM} \mathcal{L}_{eye-FSM} \\ & + \lambda_{lip-adv} \mathcal{L}_{lip-adv} + \lambda_{lip-FSM} \mathcal{L}_{lip-FSM}, \end{aligned} \quad (3)$$

where λ_{lp} , $\lambda_{comp-lp}$, $\lambda_{eye-adv}$, $\lambda_{eye-FSM}$, $\lambda_{lip-adv}$, $\lambda_{lip-FSM}$, and $\lambda_{full-adv}$ are the balancing coefficients. Here, \mathcal{L}_{L1} is the L1 loss, and \mathcal{L}_{lp} is the VGG-19 [24] based multi-scale perceptual loss [41] similar to $\mathcal{L}_{comp-lp}$. We incorporate 2-scale discriminator $\mathcal{L}_{full-adv}$ with the non-saturating loss:

$$\mathcal{L}_{full-adv} = -\log[\text{Disc}_{full}(\hat{D})], \quad (4)$$

where Disc denotes a discriminator adopted from [14]. To improve the fidelity of the facial components, we also incorporate the facial component discriminators with the feature style matching (FSM) [34],

$$\mathcal{L}_{x-adv} = -\log[\text{Disc}_x(\hat{D}_x)], \quad (5)$$

$$\mathcal{L}_{x-FSM} = \|\text{Gram}(\psi(D_x)) - \text{Gram}(\psi(\hat{D}_x))\|_1, \quad (6)$$

where $x \in \{\text{eye}, \text{lip}\}$. D_x and \hat{D}_x represent the region of interest (RoI) for the component x in the driving D and reconstruction \hat{D} , respectively. Gram is a gram matrix calculation [9] and ψ is the multi-resolution features extracted by the learned component discriminators.

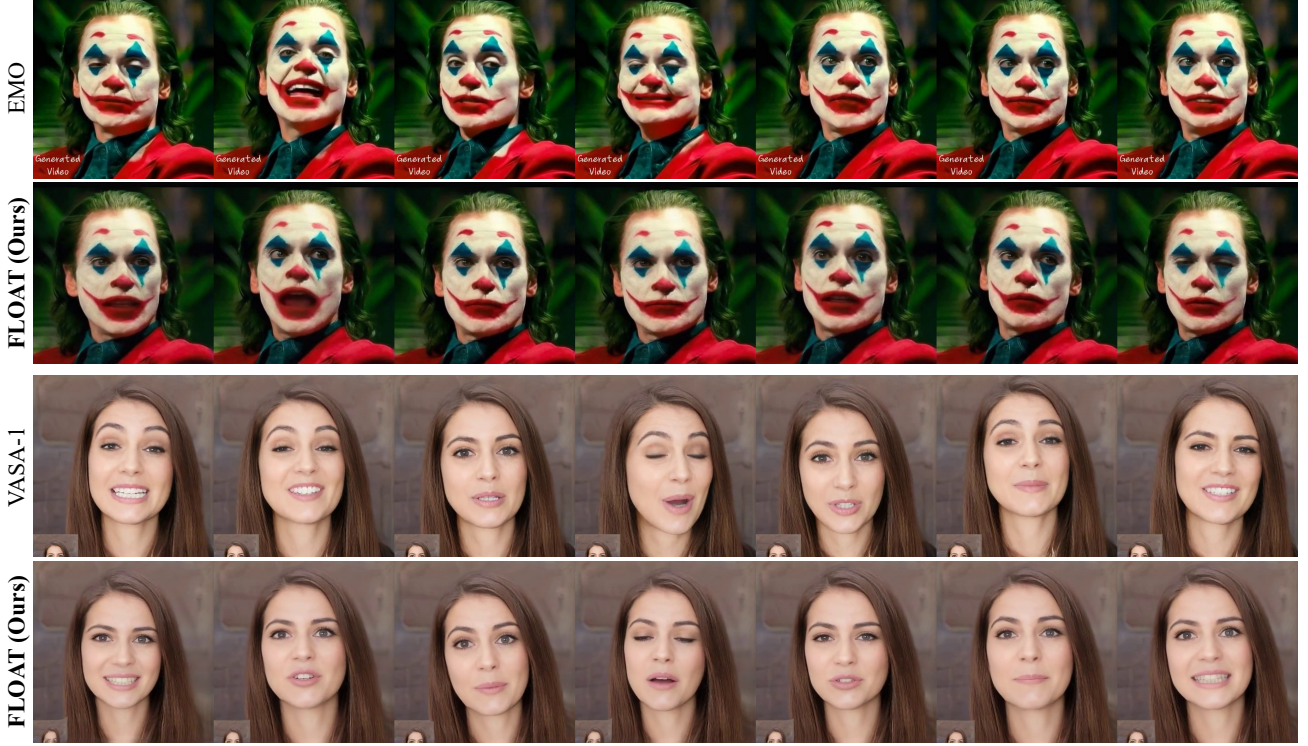


Figure 3. Comparison results with **EMO** [30] and **VASA-1** [38] based on their demonstration videos. Please note that their implementation are unavailable.

A.4. Implementation Details

We set the balancing coefficients $\lambda_{lp} = 10$, $\lambda_{comp-lp} = 100$, $\lambda_{eye-adv} = 1$, $\lambda_{eye-FSM} = 100$, $\lambda_{lip-adv} = 1$, $\lambda_{lip-FSM} = 100$, and $\lambda_{full-adv} = 1$. We employ Adam optimizer [15] with a batch size of 8 and a learning rate of $2 \cdot 10^{-4}$. Entire training takes about 9 days for 460k steps on a single NVIDIA A100 GPU.

For training our motion latent auto-encoder, we use VFHQ [36] to supplement the limited number of identities provided by HDTF [43] and RAVDESS [19]. After the same pre-processing, remaining 14,362 video clips are used for training, and 49 video clips are used for test, respectively.

B. More on FLOAT

In this section, we provide more details on FLOAT, including model, experiments, and further results.

In Fig. 9, we provide a detailed model architecture for the driving conditions c_t .

B.1. Evaluation Metrics

We provide further details of following metrics.

- **LPIPS** [41] is used to measure the perceptual similarity between reconstructed image and real image based on the pre-trained AlexNet features [16].

- **FID** [23] aims to measure the distance between the feature distributions of real and generated datasets. It is computed as:

$$\|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}), \quad (7)$$

where μ_r , Σ_r and μ_g , Σ_g are the means and covariances of the pre-trained InceptionNet [27] features from the real and generated datasets, respectively.

- **FVD** [32] is a variant of FID [23], which is used to measure the spatio-temporal consistency between the real and generated datasets by leveraging the features of pre-trained video model [2]. We compute this using 16 frames with a sliding window manner for each video.
- **CSIM** [5] measures face similarity between the two face images by computing the cosine similarity between the pre-trained ArcFace features [5] of two images.
- **E-FID** [30] aims to measure expression similarity by computing the FID score (Eq. (7)) of 3DMM expression parameters (64-dim) [6] of generated videos and real videos.
- **P-FID** aims to measure the head pose similarity by computing the FID score (Eq. (7)) of 3DMM pose parameters (6-dim) [6] of generated videos and real videos.
- **LSE-D** and **LSE-C** [20] measure lip synchronization using the pre-trained SynNet [4]. LSE-D computes the distance between the predicted audio embedding and the pre-

dicted video embedding, while LSE-C represents the confidence of synchronization.

B.2. Baselines

For non-diffusion-based methods, we compare with SadTalker [42] and EDTalk [28]. For diffusion-based methods, we compare with AniTalker [18], Hallo [37], and EchoMimic [3].

- **SadTalker** [42] employs an audio-conditional variational auto-encoder (VAE) to synthesize the head motion and eye blink in a probabilistic way.
- **EDTalk** [28] uses normalizing for audio-driven head motion generation and can separately control the lip and head motion.
- **AniTalker** [18] introduces a diffusion model to the learned motion latent space (similar to FLOAT) along with a variance adapter to improve the motion diversity. We use HuBERT audio feature-based implementation [12] for improved lip synchronization and apply default guidance scales and denoising steps of the official implementation.
- **Hallo** [37] utilizes the pre-trained StableDiffusion [22] as its image generator, incorporating a hierarchical audio attention module to separately control lip synchronization, expression, and head pose. We use default guidance scales and denoising steps provided in the official implementation.
- **EchoMimic** [3] is also StableDiffusion-based method, which leverages facial skeleton as additional driving signals. We use the default guidance scales and denoising steps provided in the official implementation.
- It is worth noting that we compare with two superior works **EMO** [30] and **VASA-1** [38] based on their demonstration videos due to their unavailable implementation. We highly recommend referring to ‘01_EMO_VASA-1_Comparison/xxx.mp4’.

B.3. More on Experiments

For evaluating our method, we use the first frame of each video clip as the source image. We use the first-order Euler method [17] as our ODE solver. We experimentally find that other ODE solvers, such as mid-point and Dopri5, do not lead to significant performance improvements.

Table 2. Ablation studies of the different NFE of ODE on HDTF [43]. FPS is computed on a single NVIDIA V100 GPU.

Ours-NFE	FID ↓	FVD ↓	E-FID ↓	LSE-D ↓	FPS ↑
Ours-2	21.785	178.831	1.542	7.559	45.22
Ours-5	21.440	164.463	1.331	7.155	44.74
Ours-10 (default)	21.100	162.052	1.229	7.290	41.37
Ours-20	21.158	164.392	1.293	7.343	38.20

Ablation on NFE In general, increasing the number of function evaluation (NFE) reduces the solution error of

ODEs. As shown in Tab. 2, even with small NFE = 2, FLOAT can achieve competitive image quality (FID) and lip synchronization (LSE-D). However, it struggles to capture consistent and expressive motions (FVD and E-FID), resulting in shaky head motion and a static expression. This is because FLOAT generates the motion in the latent space, while image fidelity is determined by the auto-encoder. We provide supplementary videos, illustrating the impact of different NFE (Number of Function Evaluations). Notably, with a small NFE of 2, the generated images exhibit good quality, but the head movements appear temporally unstable, and emotions may be exaggerated. Please refer to supplementary videos for temporal jitters of low NFE.

Table 3. Ablation studies of the audio guidance scale γ_a and the emotion guidance scale γ_e on RAVDESS [19].

Guidance scales	FID ↓	FVD ↓	E-FID ↓	LSE-D ↓
$\gamma_a=1, \gamma_e=1$	33.066	171.047	1.555	7.049
$\gamma_a=1, \gamma_e=2$	31.844	166.041	1.334	7.212
$\gamma_a=2, \gamma_e=1$ (default)	31.681	166.359	1.367	6.994
$\gamma_a=2, \gamma_e=2$	32.253	162.658	1.351	6.994

Ablation on Guidance Scales In Tab. 3, we conduct ablation studies on guidance scales: γ_a and γ_e , with the emotion intensive dataset RAVDESS [19]. Note that increasing γ_a leads to better temporal consistency (FVD) and lip synchronization quality (LSE-D). Moreover, increasing γ_e improves video consistency (FVD) and expressiveness (E-FID). This enables balanced control over emotional audio-driven talking portrait generation.

In Fig. 11, we visualize the effect of different emotion guidance scale γ_e . For this experiments, the predicted speech-to-emotion label is *disgust* with 99% probability. Notably, as increasing γ_e from 0 to 2, we can observe that emotion-related expressions and motions are enhanced.

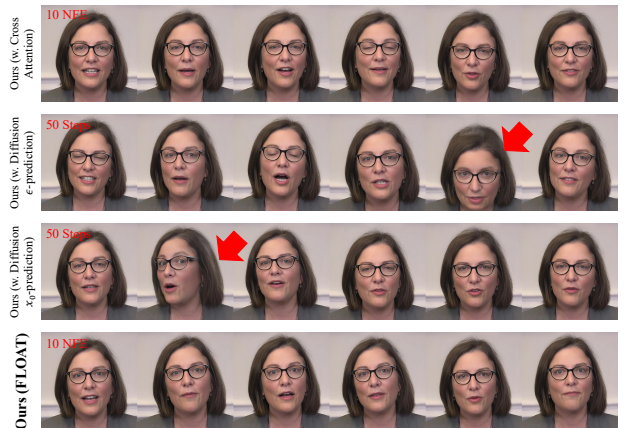


Figure 4. Ablation results on frame-wise AdaLN and flow matching. Please refer to supplementary video for notable differences.

Ablation on AdaLN and Flow Matching We conduct ab-

lation study on frame-wise AdaLN by comparing it with a cross-attention. We adopt the stand cross-attention mechanism described in [8, 26], using transformer encoder architecture for non-autoregressive sequence modeling. We use the same attention mask used in the frame-wise AdaLN, which attends to additional $2T$ adjacent frames for the l -th input latent: $[l - 2, l - 1, l, l + 1, l + 2]$.

To compare against flow matching, we implement two diffusion models with distinct parameterizations: ϵ -prediction and x_0 -prediction. For ϵ -prediction, we directly predict Gaussian noise by the noise predictor $s(\cdot; \theta)$ parameterized by θ with the following simple loss:

$$\mathcal{L}_{\text{simple, noise}}(\theta) = \|s(x_t, \mathbf{c}_t; \theta) - \epsilon\|_2^2, \quad (8)$$

where $t \sim \mathcal{U}[0, 1]$, $\epsilon \sim \mathcal{N}(0^{-L':L}, I)$, and the noise input $x_t \in \mathbb{R}^{(L'+L) \times d}$ is sampled from a forward diffusion process $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I)$ [11]. In our case, x_t is noisy motion latents at diffusion time step t , starting from $t = 0$ with $x_0 = w_{r \rightarrow D^{1:L}} \in \mathbb{R}^{(-L'+L) \times d}$.

For x_0 -prediction, we predict a clean sample x_0 , instead of noise [21], by the predictor $s(\cdot; \theta)$ with the following simple loss:

$$\mathcal{L}_{\text{simple}, x_0}(\theta) = \|s(x_t, \mathbf{c}_t; \theta) - x_0\|_2^2. \quad (9)$$

We also incorporate a velocity loss [29]:

$$\mathcal{L}_{\text{vel}, x_0}(\theta) = \|\Delta s - \Delta x_0\|_2^2, \quad (10)$$

where Δs and Δx_0 are the one-frame difference along the time-axis for s and x_0 , respectively. The total loss $\mathcal{L}_{\text{total}, x_0}(\theta)$ is

$$\mathcal{L}_{\text{total}, x_0}(\theta) = \mathcal{L}_{\text{simple}, x_0}(\theta) + \mathcal{L}_{\text{vel}, x_0}(\theta). \quad (11)$$

For reverse process, we use the DDIM [25] sampler with 50 denoising steps.

In our implementation, both ϵ -prediction and x_0 -prediction achieve the best results with guidance scales $\gamma_a = \gamma_e = 1$ (default). In Fig. 4, Fig. 12 and Fig. 13, we provide qualitative comparisons between these approaches and FLOAT. Notably, the cross-attention exhibits less diverse head motions compared to FLOAT, while diffusion-based approaches struggle to generate temporally stable lip and head motion, often resulting in out-of-sync movements or motion artifacts.

C. Additional Results

C.1. Additional Comparison Results

We provide additional comparison results with baselines in Fig. 15, Fig. 16, and Fig. 17.

C.2. Out-of-distribution (OOD) Results

In Fig. 10 and Fig. 11, we present additional out-of-distribution results, including paintings, non-English speech, and singing.

C.3. User Study

Table 4. Mean opinion score (MOS) study results with 95% confidence interval. The score ranges in 1 to 5. The best result for each metric is in **bold**.

Method	Lip Sync Accuracy	Natural Head Motion	Teeth Clarity	Natural Emotion	Overall Visual Quality
SadTalker [42]	2.20 \pm 0.35	2.03 \pm 0.26	1.53 \pm 0.19	1.80 \pm 0.28	1.97 \pm 0.23
EdTalk [28]	2.50 \pm 0.34	2.60 \pm 0.28	1.17 \pm 0.17	2.07 \pm 0.36	1.83 \pm 0.27
AniTalker [18]	2.70 \pm 0.31	3.00 \pm 0.30	2.13 \pm 0.27	3.17 \pm 0.27	2.63 \pm 0.26
Hallo [37]	3.30 \pm 0.32	2.73 \pm 0.35	2.23 \pm 0.27	2.67 \pm 0.35	2.27 \pm 0.33
EchoMimic [3]	2.67 \pm 0.37	3.07 \pm 0.30	2.20 \pm 0.34	2.50 \pm 0.37	2.70 \pm 0.36
FLOAT (Ours)	3.93 \pm 0.21	3.57 \pm 0.33	4.13 \pm 0.27	3.77 \pm 0.30	3.87 \pm 0.30

In Tab. 4, we conduct a mean opinion score (MOS) based user study to compare the perceptual quality of each method (e.g., teeth clarity and naturalness of emotion). We generate 6 videos by using the baselines and FLOAT, and ask 15 participants to evaluate each generated video with five evaluation factors in the range of 1 to 5. As shown in Tab. 4, FLOAT outperforms the baselines.

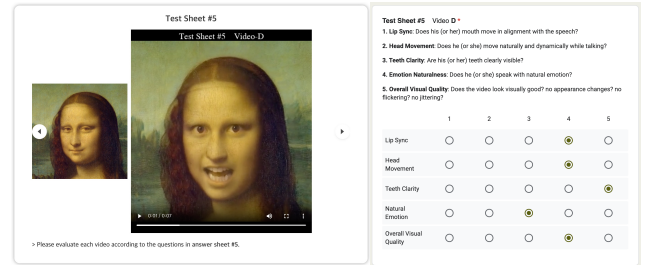


Figure 5. Example of user study interface. (Left) Test Sheet; (Right) Answer Sheet. Participants were asked to evaluate 5 questions for each video (total 180 videos).

In Fig. 5, we provide an example of test and answer sheet used of the user study. We asked 15 participants to evaluate five questions for each generated video produced by the baselines and FLOAT. Consequently, each participant scores total 180 questions, with responses ranged from 1 to 5. Additionally, we include the supplementary videos used in the user study.

C.4. Video Results

We include video results to further illustrate the performance of our method, including emotion redirection, additional driving conditions, and OOD results. Please refer to provided videos.

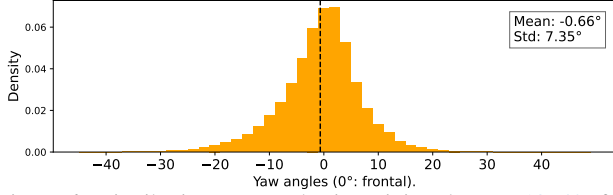


Figure 6. Distribution yaw angles in training dataset [19, 43] for FLOAT.

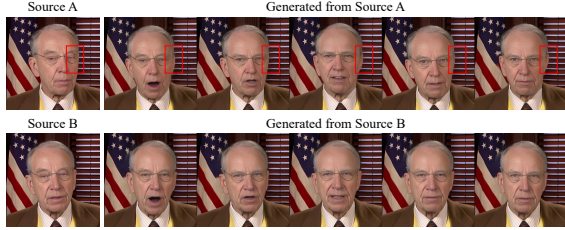


Figure 7. Failure case of FLOAT. It often struggles to handle non-frontal faces and accessories, such as glasses. Please refer to supplementary video.

D. Discussion

Ethical Consideration This work aims to advance virtual avatar generation. However, as it can generate realistic talking portrait only from a single image and audio, we considerably recognize the potential for misuse, such as deepfake creation. Attaching watermarks to generated videos and carefully restricted license can mitigate this issues. Additionally, we encourage researchers in deepfake detection to use our results as data to improve detection tools.

Limitation and Further Work While our method can generate realistic talking portrait video from a single source image and a driving audio, it has several limitations.

First, our method cannot generate more vivid and nuanced emotional talking motion. This is because the speech-driven emotion labels are restricted to seven basic emotions, making it challenging to capture more nuanced emotions like *shyness*. We believe this limitation can be addressed by incorporating textual cues (e.g., “gazing forward with a shyness”), an idea we plan to explore in future work. Moreover, any other approaches to enhance the naturalness of talking motion are key directions for our future work.

Second, we aim to build our method solely upon high-definition open-source datasets. Since the training datasets are biased toward frontal head angles [19, 43], the generated results also exhibit a similar bias, often producing suboptimal results for non-frontal (e.g., $|\text{yaw angle}| \geq 20^\circ$) source images or images with notable accessories. This is partially because the head pose distribution of our training data as shown in Fig. 6. Although we investigated other existing high-definite face video datasets, such as MEAD [33] and CelebV-Text [40], we found limitations in their suitability. MEAD [33] contains minimal head motion and a limited

number of identities, while CelebV-Text [40] is not organized for audio-driven talking portrait, containing out-of-sync audio and significant background inconsistencies.

This limitations can be mitigated by introducing carefully curated external data, as demonstrated by other concurrent methods [10, 13, 30, 37, 38], or by incorporating multi-view supervision [31] when training our motion latent auto-encoder. We provide examples of failure case in Fig. 7 and supplementary video.

Acknowledgment The source images and audio used in this paper are taken from other talking portrait generation methods [3, 30, 37, 38, 42]. We sincerely thank the authors of these works for their valuable contributions. Note that the individuals depicted in our source images and the speech generated in our experiments are not associated with the actual persons they represent.

References

- [1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 2
- [2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3
- [3] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 4, 5, 6
- [4] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision*, pages 251–263, 2016. 3
- [5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 3
- [6] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3
- [7] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2
- [8] 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 (CVPR)*, pages 18770–18780, 2022. 5
- [9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 2
- [10] Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023. 6
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5
- [12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460, 2021. 4
- [13] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 6
- [14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 2, 9
- [15] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 3
- [17] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4
- [18] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding. *arXiv preprint arXiv:2405.03121*, 2024. 4, 5
- [19] Steven R Livingstone and Frank A Russo. The ryer-son audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 2, 3, 4, 6
- [20] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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*, pages 484–492, 2020. 3
- [21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 5
- [22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 4
- [23] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, 2020. Version 0.3.0. 3
- [24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2
- [25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5
- [26] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024. 5
- [27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on*

- computer vision and pattern recognition, pages 1–9, 2015. 3
- [28] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2025. 4, 5
- [29] 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. 5
- [30] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 3, 4, 6
- [31] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 6
- [32] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 3
- [33] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 6
- [34] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9168–9178, 2021. 1, 2
- [35] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 1, 2, 9
- [36] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. 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)*, pages 657–666, 2022. 2, 3
- [37] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 4, 5, 6
- [38] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 3, 4, 6
- [39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 2
- [40] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023. 6
- [41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 1, 2, 3
- [42] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8652–8661, 2023. 4, 5, 6
- [43] 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 (CVPR)*, pages 3661–3670, 2021. 2, 3, 4, 6

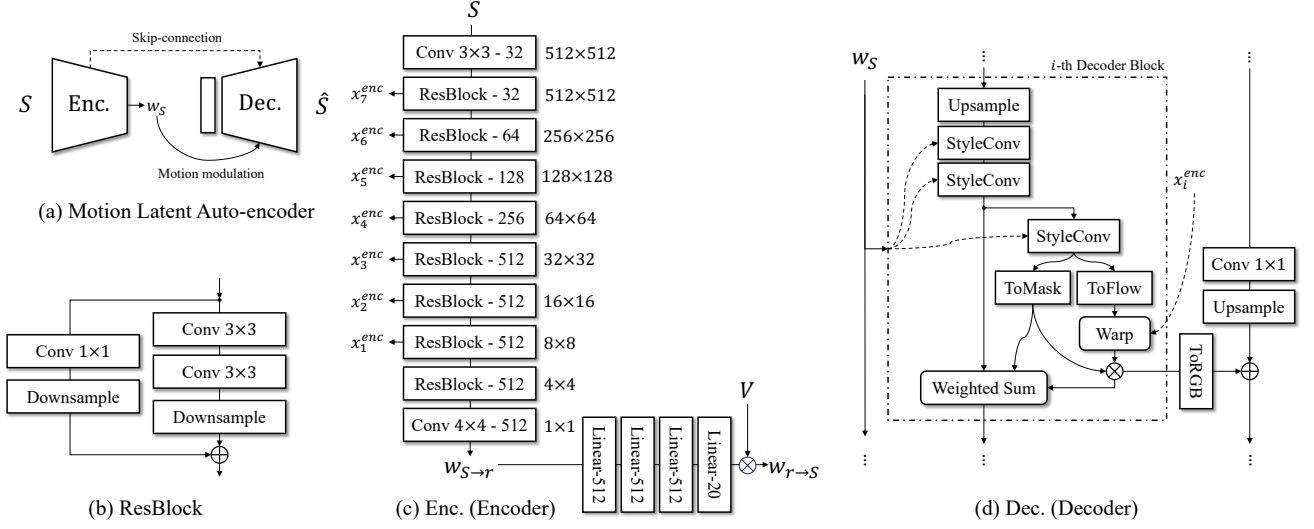


Figure 8. Detailed Model architecture of our motion latent auto-encoder. The notations are adopted from LIA [35] and StyleGAN2 [14].

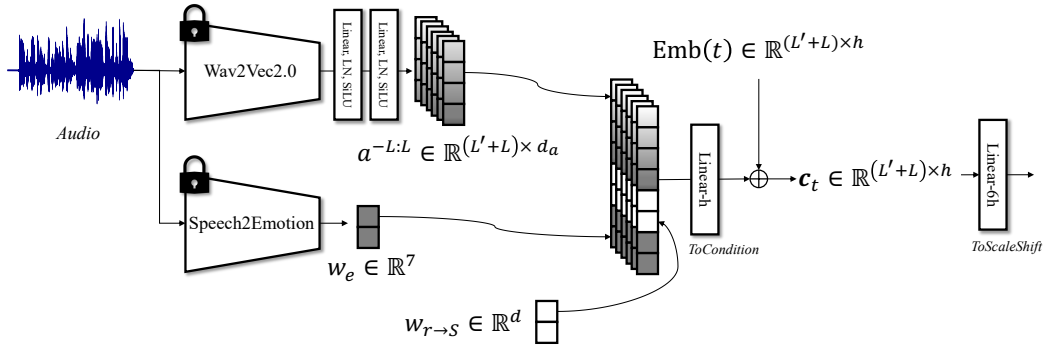


Figure 9. Detailed model architecture for constructing the driving conditions $\mathbf{c}_t \in \mathbb{R}^{(L'+L) \times h}$ in FLOAT.

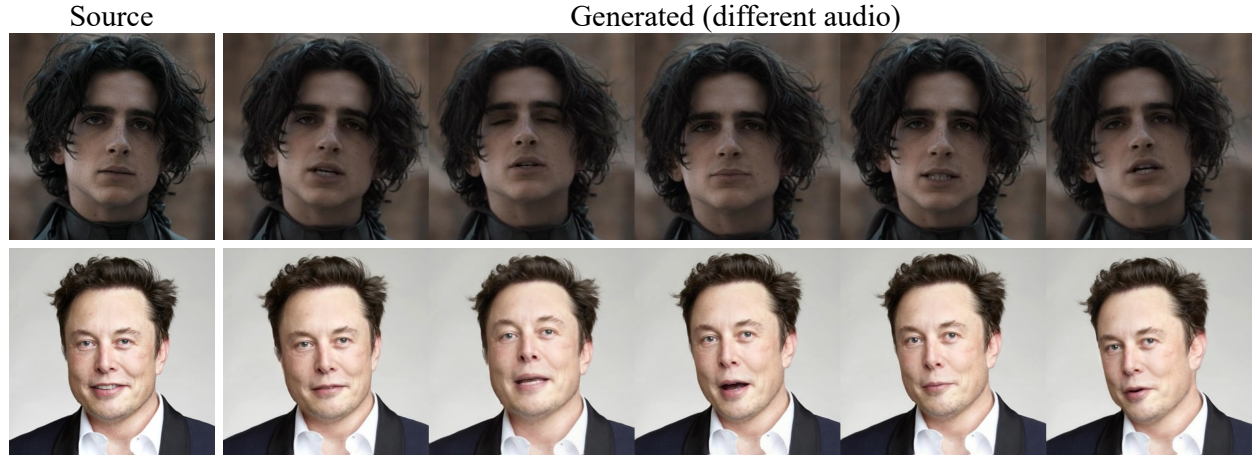


Figure 10. Out-of-distribution results. The first row shows the result for *Chinese* audio, and the second row shows the result for *singing* audio. Please refer to supplementary video.



Figure 11. Ablation on emotion guidance scale γ_e . The predicted speech-to-emotion label is *disgust* of 99.99%. Please refer to supplementary video.



Figure 12. Ablation results on frame-wise AdaLN and flow matching. Please refer to supplementary video.



Figure 13. Ablation results on frame-wise AdaLN and flow matching. Please refer to supplementary video.



Figure 14. Ablation results on frame-wise AdaLN and flow matching. Please refer to supplementary video.



Figure 15. Qualitative comparison results with state-of-the-art methods. Please refer to supplementary video.



Figure 16. Qualitative comparison results with state-of-the-art methods. Please refer to supplementary video.



Figure 17. Qualitative comparison results with state-of-the-art methods. Please refer to supplementary video.