arXiv:2312.03775v1 [cs.CV] 6 Dec 2023

FAAC: Facial Animation Generation with Anchor Frame and Conditional Control for Superior Fidelity and Editability

Linze Li Sunqi Fan Hengjun Pu Zhaodong Bing Yao Tang Tianzhu Ye
Tong Yang Liangyu Chen Jiajun Liang†

Figure 1. Our method provides three style high-fidelity and edited face animation: training-free FACC, training-based FACC and FACC with conditional control. Compared with Animatediff[6] or its motion module in different resolutions, FACC shows priority on the facial fidelity, text-to-image editability and video motion. MM stands for motion module, 64, 32, and 16 represent resolutions of 64x64, 32x32 and 16x16 respectively. T2I wo MM is the theoretical performance ceiling for facial fidelity and editing abality.

Abstract

Over recent years, diffusion models have facilitated significant advancements in video generation. Yet, the creation of face-related videos still confronts issues such as low facial fidelity, lack of frame consistency, limited editability and uncontrollable human poses. To address these challenges, we introduce a facial animation generation method that enhances both face identity fidelity and editing capabilities while ensuring frame consistency. This approach incorporates the concept of an anchor frame to counteract the degradation of generative ability in original text-to-image models when incorporating a motion module. We propose two strategies towards this objective: training-free and training-based anchor frame methods. Our method's efficacy has been validated on multiple representative DreamBooth and LoRA models, delivering substantial improvements over the original outcomes in terms of facial fidelity, text-to-image editability, and video motion. Moreover, we introduce conditional control using a 3D parametric face model to capture accurate facial movements and expressions. This solution augments the creative possibilities for facial animation generation through the integration of multiple control signals. For additional samples, please visit our anonymous project page.

1. Introduction

With the breakthrough development of deep learning and artificial intelligence, video generation technology has made significant progress in recent years. In particular, methods based on diffusion models have brought new opportunities to the field of video generation. Among these, facial anima-

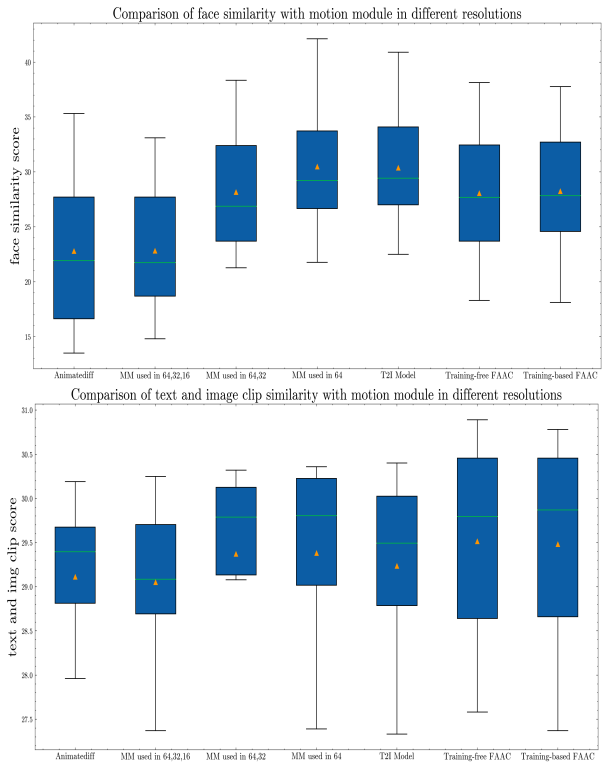

Figure 2. Comparison of face similarity and text to image clip similarity with motion module in different resolutions and our methods

tion generation, as an important branch of computer vision research, has attracted widespread attention due to its rich application scenarios, such as film production, virtual reality, and social media. High-quality facial animation generation needs to satisfy requirements of realism, expressive details, and ease of editing and control.

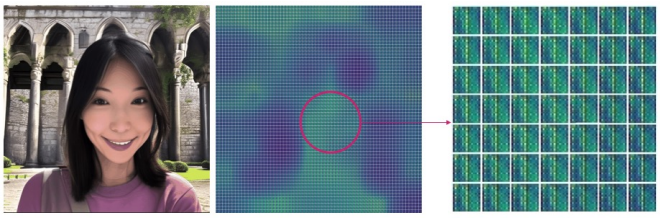

Figure 3. Visualization for Observation 2: Averaged attention maps from Temporal attention Module of each pixel. The model always attends to features of the same frame, causing the activation of a single column in the heat map.

Despite this, a series of challenges still remain in high-quality facial animation generation, including realism, fidelity, and expressive detail. For example, although traditional face-swapping methods (such as Deepfakes[15]) can achieve a certain degree of realism, they are limited in terms of expression details and diversity, and they often struggle to maintain the fidelity of face identity with occlusions and pose changes. Also, their processing scope is limited to the facial area, making it impossible to effectively edit non-facial regions, thus reducing the creativity and editability of facial animation generation. Considering about creativity and editing of images or videos, the diffusion model, which has been particularly popular recently and has strong performance, comes to our mind. To our knowledge, the most popular and outstanding face or portrait video generation model based on the diffusion model is perhaps AnimateDiff[6]. The authors proposed a framework for extending any customized text-to-image model for animation generation, which can generate corresponding animation clips while maintaining the image style of the original customized model.

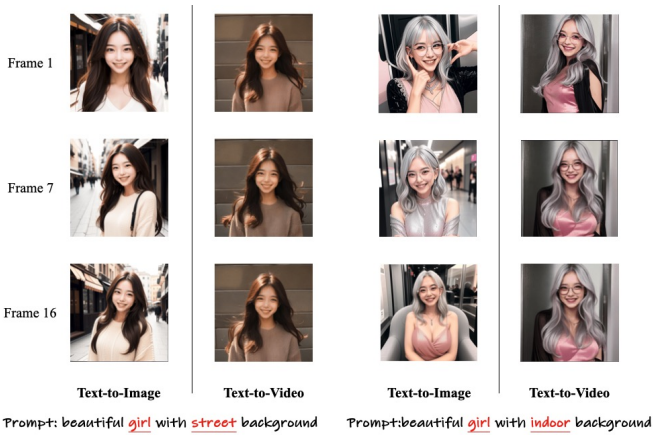

Figure 4. Visualization for Observation 3: Difference between T2I and T2V generated samples from same Gaussian noise start. Training for coherent frames will result in a simpler background and details.

We have three observations about the motion module of Animatediff[6] which inserted in the Stable Diffusion Model for video generation. **Observation 1: You can't have your cake and eat it too. Motion Module disrupts the generative ability of the original text-to-image model.** AnimateDiff[6] inserts a motion modeling module into each resolution of the text-to-image model. We found that the deep motion module disrupts the generative ability of the original text-to-image model. As shown in Figure 1, as the layers deepen, the facial fidelity and editability of the model deteriorate, especially at the 16x16 and 8x8 stages. Figure 2 further quantifies the facial fidelity and editability of Animatediff, demonstrating a significant improvement when the deep motion model is discarded in the low resolution of latent. On the other hand, as motion module insert in the low resolution, the inter-frame continuity of video generation gradually strengthens. This raises the

question: Is it possible to maximize the facial fidelity and editability to approach that of the text-to-image model without losing inter-frame continuity? **Observation 2: Keep an eye on the person ahead, makes the line move fast instead. Motion Module tends to align with the middle frame in the early stages of denoising.** As shown in Figure 3, we performed a visualization analysis of the temporal attention score maps and found that the model aligns with the middle frame already at timesteps 15. This reminds us that explicit modeling consistency to the anchor frame may help the model work better. **Observation 3: The shortest distance between two points is a straight line. Training for consistency frames makes generalized background simple and boring.** In order to meet the consistency constraints for frames, the model has a certain probability of favoring simpler backgrounds. As shown in Figure 4, we provided the prompt "with street background" and "with indoor background" independently, the original T2I model tends to generate complex street scenes and indoor scenes, while AnimateDiff[6] has an obviously higher probability of generating a simple wall and pure background.

In this work, we propose a novel facial animation generation scheme aimed at achieving realistic, smooth, high-fidelity, and richly detailed facial animations while enhancing the generation and editing capabilities of non-facial regions. We introduce the concept of an Anchor Frame to avoid the insertion of the motion model damaging the facial fidelity and editability of the text-to-image model based on observation 1 and observation 2. We provide both a training-free Anchor Frame inference method and a training-based Anchor Frame inference method. Both the training-free method and training-based method can improve both facial fidelity and editing capabilities by modeling the consistency with the anchor frame in the training process. Besides, in terms of facial fidelity and details, we introduce a conditional control using a 3D parametric face model to make the capture of facial movements and expressions more accurate.

To validate the effectiveness of the proposed method, we evaluate our AnimateDiff[6] on several representative DreamBooth and LoRA models about realistic photographs. Whether it's facial fidelity, text-to-image editability, or video motion, we have significantly improved compared to the original results. Additionally, we support combined conditional control generation and long-sequence video generation options, thus providing a broader creative space for facial animation generation.

The main contributions of this paper are as follows:

- We propose a novel facial animation generation method that aims to generate realistic, smooth, high-fidelity, and richly detailed facial animation while enhancing the generation and editing capabilities of non-facial regions. We introduce the training-free and training-based anchor frame method, which both counteract potential issues where the motion model might damage the facial fidelity and editability of the text-to-image model.
- We introduce a conditional control using a 3D parametric face model, making the capture of facial movements and expressions more accurate. Additionally, we support combined conditional control generation and long-sequence video generation options, thereby providing a broader creative space for facial animation generation.
- We validate the effectiveness of our proposed method on multiple representative DreamBooth and LoRA models, and we have made significant improvements compared to the original results in terms of facial fidelity, text-to-image editability, or video motion.

2. Related work

2.1. Conditioned Video Generation with Diffusion Models

Video Diffusion Models (VDM)[8] first extended the DDPM[7] models initially used for text-to-image generation, utilizing a factorized space-time U-Net to execute temporal attention based on the Text-to-Image (T2I) model. Then it becomes common to extend a T2I model with temporal structures for video generation. Animatediff[6] inserts motion modules into the U-Net to learn appropriate motion priors given textual descriptions. In addition, personalized generation can be achieved in Animatediff[6] by modifying the T2I model from base model (e.g., Stable Diffusion[19]) to personalized models, such as DreamBooth[20] (utilizing a rare string) and Lora[9] (fine-tuning weights' residuals via low-rank decomposition).

However, text descriptions often struggle to accurately represent complex motion. To enhance motion control within videos, some methods incorporate pose (e.g., Follow Your Pose[12]) or trajectories (e.g., DragNUWA[26]) to facilitate continuous video control. As a variety of conditions emerge, several works aim to develop frameworks that accommodate these diverse guidance. VideoComposer[23] suggests categorizing conditions into three types (textual, spatial, and temporal) and employing condition fusion to leverage different control signals for collective guidance. Some works on image generation have also accomplished similar tasks. ControlNet[29] employs a trainable copy of the initial neural network block with zero convolution layers, which receives extra conditions as input and adds the output to the original result. T2I-Adapter[14] designs a simple adapter to extract multi-scale condition features. These methods can assist to synthesize videos since a T2I model can be extended to a Text-to-Video (T2V) model as previously mentioned.

2.2. Face Generation with Diffusion Model

Recently, diffusion models have been gradually applied to the highly practical domain of the human face. DiffSwap[30] firstly leverages the conditional inpaiting capability of diffusion models to perform face swapping task.To generate customized human portrait, MagiCapture[11] follows DreamBooth[20] to allocate special tokens for both source face and reference style, achieving disentanglement between face and background. Recent efforts have been increasingly focused on developing methodologies for generating animatable 3D-aware human facial models. Text2Control3D[10] leverages diffusion models with controlnet[29] to generate multi-view face images, and then use these images to construct neural implicit field. Following DreamFusion[17], FDNeRF[27] designs a diffusion loss to optimize the latent code representing a face, achieving a controllable prompt-driven face editing result. DreamFace[28] generates animatable 3D face through three key steps: geometry generation, physically-based texture diffusion, and animation empowerment.

3. Method

As shown in Figure 5, our method includes a frozen Text-to-Image diffusion model (T2I model)., a Temporal attention Module, and a Ghost Module. We first give a brief introduction to video generation from T2I models in 3.1. Training and inference schemes of our method FAAC are further discussed in 3.2 and 3.3.

3.1. Preliminary:Video Generation from Text-to-Image diffusion Models

Diffusion models are generative models that generate samples that fit the real image distribution, recovering from totally random noise by a reversed denoising process. The forward process of diffusion models can be described as: $q_t(x_t|x_0) = \mathcal{N}(x_t|\alpha(t)x_0, \beta(t)I)$, where x_0 is a sample from original image distribution and $\alpha(t), \beta(t)$ are the noise scheduler weight. By estimating the noise and reversing the process, we can generate a novel sample by:

$$\begin{aligned} p_\theta(x_{0:T}) &= p_(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \\ p_\theta(x_{t-1}|x_t) &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)) \end{aligned} \tag{1}$$

where $\mu_\theta, \sigma_\theta$ is estimated by neural networks θ. The training objective of the diffusion model is the simple reconstruction loss:

$$L_{simple}(\theta) = ||\epsilon_\theta(x_t, t) - \epsilon||^2 \tag{2}$$

where ϵ is Gaussian random noise injected into a noisy sample x_t by noise scheduler. Due to the success of diffusion models in capturing realistic images and video priors, there have been some excellent work of Text-to-Image diffusion models such as Latent Diffusion Models (LDM)[19]. LDM works on a lower resolution of latents so that the time efficiency and generation fidelity are well-balanced. There are two components in LDM: an autoencoder which encodes images and decodes the generated latents into images, and a normal diffusion model (modified U-Net architecture) which functions on noisy latents.

The idea of turning image diffusion models into video diffusion models is very natural. AnimateDiff [6] injects Temporal attention Module into the original LDM U-Net and trains the Temporal attention Module (cross-frame pixel-wise attention) on video datasets to generate text-to-video samples. It multiplexs the original T2I model's weight to process each single frame and uses the Temporal attention Module to get motion priors and objective consistency across frames. Our method followed AnimateDiff's architecture to use the Temporal attention Module.

3.2. Anchor Frame Motion Training

To address the disadvantages brought by the Temporal attention Module in **Observation 1** and **Observation 3**, we propose Anchor Frame Motion Training to solve the problem.

During training, we randomly choose a frame across all frames as Anchor Frame for each video batch. We annotate the anchor frame index as k. the noisy latent of Anchor Frame is sent to the frozen T2I model and the Ghost Module. The Ghost Module does not modify any features of Anchor Frame so the output of predicted noise is the same as the output of the original frozen T2I model. The only function of Ghost Module is to send features of Anchor Frame to attend with features of other frames in the Temporal attention Module. The noisy latents of other frames are all sent to the frozen T2I model and Temporal attention Module. We only fine-tune the Temporal attention Module during training. The temporal attention can be expressed by:

$$\begin{aligned} z_i &= Attention(Q_i K^T)V, i \neq k \\ Q_i &= W^Q z_i, K = W^K z, V = W^V z \\ z &= Cat(z_1, z_2, ..., z_n) \end{aligned} \tag{3}$$

where z_i is the feature of the i-th frame. The feature z_k of Anchor Frame is not changed during the denoising process. Therefore it can fully utilize the editability and the fidelity of the T2I model and will not be affected by the performance degradation brought about by the motion module.

Due to the domain gap between training data for T2I models and Temporal attention Module (one is images, which we can hardly reach, and one is videos), cumulative errors may occur in the denoising process if the training targets of other frames during training are based on data

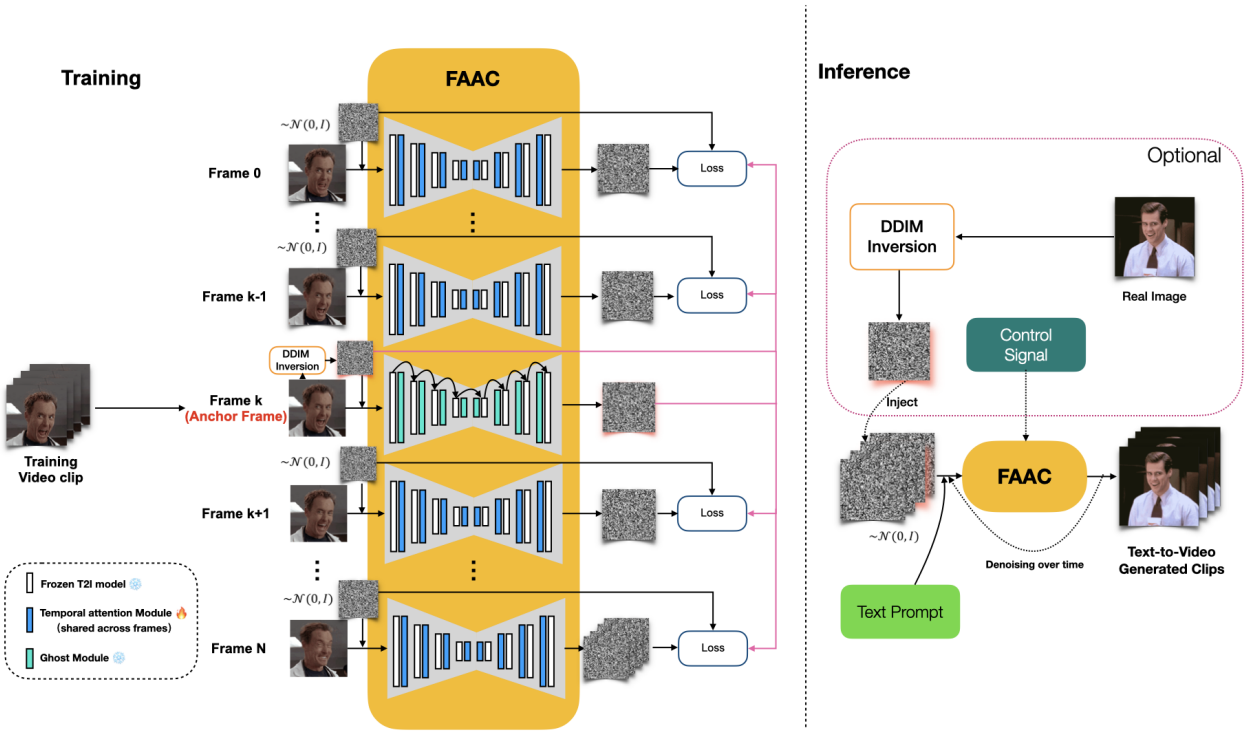

Figure 5. **Overview.** We propose a training method and an inference method for facial video generation. During training, we freeze the T2I model and train the Temporal attention Module with the help of our Anchor Frame mechanism. During inference, coherent video samples can be generated simply with our Anchor Frame Inference to get the same editability and fidelity of the original T2I model by the training-free or training-based approach.

from the video training set, but the anchor frame is not constrained. This could lead to a decrease in fidelity and abrupt alterations in the generated videos.

To tackle this problem, we replace random noise with estimated noise from DDIM Inversion for the Anchor Frame. Therefore the predicted x_0 for Anchor Frame will be the nearly same as the original frame in the training video. The gap between the two domains can be effectively minimized by employing this approach, thereby facilitating the training of more coherent models.

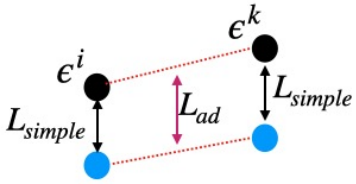

Figure 6. Graph of loss concept. Instead of learning to predict sampled random noise ϵ, the loss will push the model to learn the difference between ϵ_i and ϵ_i.

Training Objective. Besides replacing the random noise with estimated noise from DDIM inversion, we redesigned the training loss (called Anchor Difference Loss) to diminish the gap instead of the simple reconstruction loss:

$$L_{ad} = \sum_{i=1, i\neq k}^{n} ||(\epsilon_\theta(x_t^i, t) - \epsilon_\theta(x_t^k, t)) - (\epsilon^i - \epsilon^k)||^2 / n \tag{4}$$

where $\epsilon_\theta(x_t^i, t$ is the estimated noise from model of frame i in at time step t and ϵ^i is the sampled Gaussian noise of frame i.

The Anchor Difference Loss can decrease the latent gap between Anchor Frame and other frames, preserving the fidelity of faces. And the total loss of our training is $L = L_{simple} + \lambda L_{ad}$. We set λ as 1.0 for our experiment.

3.3. Anchor Frame Inference

The inference pipeline is almost the same as the training procedure, as the anchor frame is generated only from the T2I model. The original x_T of Anchor Frame can also be generated from any realistic image using DDIM Inversion, making our model suitable for animating any generated images or real images.

Training-free. We also discovered that the Temporal attention Module does not need to be fine-tuned to get quality results in our experiments. Therefore our inference method

can also be training-free, which is a plug-and-play tool for facial video generation.

Control signals. Our method is also compatible with control signals to generate facial videos with certain facial landmarks, rendered face images from 3D Morphable Models, canny images, etc. See details in Section 4.2.

4. Experiment

In the experimental section, we firstly present the details of our training and evaluation in Section 4.1. Subsequently, in Section 4.2, we elucidate how we integrate FACC with controllable generation through the utilization of 3D Morphable Model (3DMM)[1]. Following this, we discuss the qualitative and quantitative results of our method in Section 4.3, 4.4.

4.1. Experimental Detail

Training. We use Stable Diffusion v1.5 as our base text-to-image model to train the controllable modeling module of expression and pose and the motion module. We use 1000+ portrait videos as our dataset, the controllable modeling module was trained from scratch while the motion modeling module was fine-tuned with Animatediff V2 motion module. The video clips in the dataset are sampled at the stride of 1, then resized and center-cropped to the resolution of 512 × 512, the length of the video clips for training is set to 16 frames.

Evaluations. To confirm the efficacy and broad applicability of our approach, We use a diverse range of LoRA[9] collected from Civitai or trained by ourseleves, encompassing various genders, ages, and ethnicities. To enhance realism, following [6], we also colleted a great number of realistic photography style DreamBooth[20] from Civitai. Our pipeline intergrates the base text-to-image model, motion module, stylized DreamBooth, and personalized LoRA, enabling the generation of highly realistic facial animations for the LoRA characters.

4.2. FACC With Control

Based on our previous observations, Animatediff tends to prioritize ensuring inter-frame consistency in generated videos rather than inducing intricate motion. As a result, the generated video clips frequently display minimal movement in human actions, accompanied by negligible or no changes in facial expressions. The animation seems like a camera movement, and even common actions like blinking and smiling are infrequently observed.

In our pursuit of generating facial videos with more substantial motion, we leverage the power of T2I-adapter[13]. Similar to ControlNet[29], T2I-adapter is a commonly used method for controllable image generation given various conditions, such as pose, sketch, color, etc. Specifically, T2I-adapter utilizes a trainable Adapter structure to align internal knowledge in T2I models and external control signals.

The central focus of our investigation revolves around identifying a signal conducive to controlling the generation of facial animations. After careful deliberation and empirical exploration, our choice coalesces around the use of 2D rendering images obtained from a 3D Morphable Model (3DMM)[1, 3, 5] as a conditional signal to control the generation of facial animations. Compared to facial landmarks, face parsing, and canny maps on human faces, using 2D rendering images of a 3D Morphable Model as condition offers several advanteages (demonstrated in Figure 7).

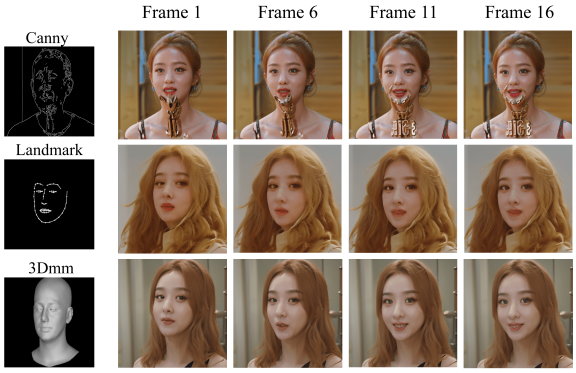

Figure 7. **The generation effects under different control signals.** It can be seen that the use of canny may result in disorderly lines that may not be correctly processed by Diffusion, thus failing to generate a reasonable facial video (the girl's chin in the first row). When generating using landmarks, it is sometimes challenging to accurately control the signal actions. Simultaneously, if the facial shape generated by the LoRA character differs significantly from that in the control signal, canny and landmark signal may lead to a decrease in fidelity. It can be observed that the use of 3DMM for generation has shown improvements in both fidelity and control accuracy of facial expressions.

Most importantly, we can recombine facial features with the assit of 3DMM. The 3DMM model enables the extraction of high-dimensional and fine-grained features such as pose, shape, expression, texture, and identity from a single image, condensed them into low-dimensional latent codes. By incorporating external features, such as the pose and expression of another individual, with the internal features of the facial shape, texture, and identity of the person we aim to generate, we enhance fidelity. Moreover, we can simultaneously capture facial features from multiple individuals. For instance, by incorporating the expression from individual A, the pose from individual B, and the facial shape and texture of the target-generated person. Figure 8 illustrates an example.

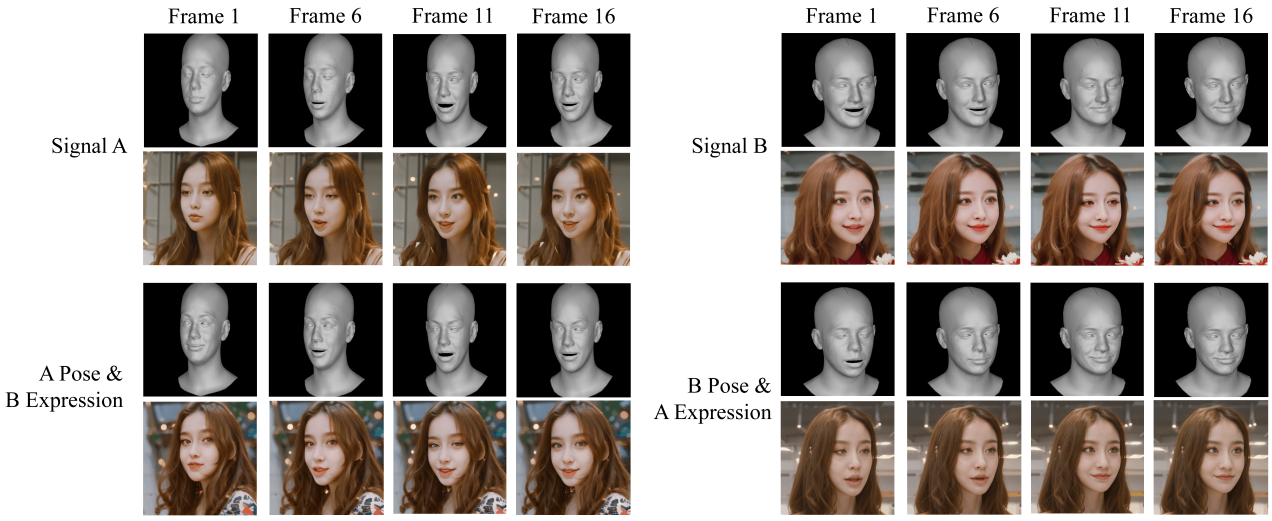

Figure 8. **Example of pose and expression recombination.** It is worth noting that the pose feature also includes mouth features in 3DMM. Therefore, the party providing the pose feature determines the control signal for the mouth.

4.3. Qualitative Results

We conduct comprehensive evaluation under the same setting to fairly compare the qualitative results between our method and the AnimateDiff[6] baseline. We collected several typical and representative examples, as demonstrated in Figure 9. It is worth noting the following three aspects:

- Our approach is more in line with the prompt compared to the baseline. Descriptions incongruent with baseline samples are highlighted in red.
- Our method exhibits higher fidelity than the baseline. Under the same LoRA conditions, the facial animations generated by our approach more closely resemble the LoRA character.
- Compared to the baseline, our approach exhibits a greater range of motion. If control signals are incorporated, our samples can generate larger ranges of motion.

4.4. Quantitative Results

In order to comprehensively assess the fidelity, editability, and the overall quality of the generated facial animations, we use the following 3 kinds of metrics for qualitative evaluation.

Face Similarity Score: we leverage ArcFace[4] to assess the fidelity. ArcFace is a deep face recognition network which can extract representative face features through an Additive Angular Margin Loss. We employed ArcFace to extract facial embeddings from both the generated facial animations and authentic photographs of the LoRA character. The cosine similarity was computed to serve as the Face Similarity Score.

CLIP Score: we utilize the CLIP to assess the text-to-image editability. Specifically, we compute the CLIP[18] similarity between the prompt input of the diffusion model and the generated animation images to examine whether the prompt effectively controls and edits the generation of video content.

Fréchet Video Distance: following prior works[16, 21, 25], we utilize the Fréchet Video Distance (FVD) [22] to assess the overall quality of our generated video clips. FVD initially employs the pretrained I3D video classification network[2] to extract feature representations from both real and synthesized videos. Subsequently, it computes the Fréchet distance between the distributions of features from real and synthesized videos. Following [21, 24], we utilize CelebV-HQ[31] dataset as our real benchmark. And we use our approach and a baseline to synthesis a large number of videos under the same set of setting (prompts, DreamBooth, and LoRA) for fair comparison.

Table 1. Quantitative Comparison

Method	**Fidelity** (↑)	**Editability** (↑)	**FVD** (↓)
AnimateDiff	22.79	29.11	595.45
training-free FACC	28.04	29.48	**504.15**
training-based FACC	**28.24**	**29.51**	523.27

Under the above three metrics, we compare our methods with AnimateDiff[6]. We observed a significant improvement in fidelity, editability, and overall video quality for both training-free and training-based FACC compared to AnimateDiff. While the training process may hurt the FVD metric, it concurrently enhances fidelity and editability.

Figure 9. **Qualitative Comparison.** Video sequence produced by our method (the second line in each unit) more closely align with the textual description and the LoRA character compared with the baseline (the first line in each unit).

5. Limitations and Future Works

Despite the promising results achieved by our method, we also recognize several limitations and potential avenues for future exploration.

In our anchor frame approach, we observed that the image quality of non-anchor frames is lesser compared to anchor frame. The denoising process resembles a chasing problem: despite striving to match the anchor frame, it never reaches perfection. Furthermore, in Observation 2, we mentioned that inserting a motion model at a deep level into the T2I model results in loss of the original facial fidelity and editability provided by the T2I model. However, our current approach has not yet capitalized on this insight, marking an area for future investigation.

Another significant aspect of our research accentuates the importance of conditional control using a 3D parametric face model. This feature enables us to capture facial movements and expressions more accurately, contributing significantly to our method's performance. However, we found that introducing control signals, such as 3DMM, harms the original fidelity of the T2I model—a challenge worth continuing research. Moreover, how to generate control signals from text is a research topic in anticipation of accomplishing comprehensive fine-motion video generation merely using text.

Additionally, our current facial fidelity is entirely sourced from the LoRA model. Exploring other methods to generate videos of specific ID faces is also a direction for our future work.

6. Conclusion

In this work, we proposed a novel facial animation generation approach, using the potential of diffusion models enriched with our newly introduced anchor frames and conditional control. Our technique effectively handles previously observed challenges in fidelity and editability when incorporating motion dynamics into Text-to-Image (T2I) models.

The integration of a 3D parametric face model added to our methodology by providing a more accurate capture of facial movements and expressions, contributing to the dynamic realism and the 3D-consistency of the generated animations. Furthermore, our method's capacity for combined conditional control generation opened up new possibilities for creative applications.

Our experimental results validated our method's effectiveness on multiple representative DreamBooth and LoRA models, showing significant improvements in facial fidelity, text-to-image editability, and video motion compared to existing solutions.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 6

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 7

[3] Radek Daněček, Michael J. Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20279–20290, 2022. 6

[4] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. 7

[5] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images, 2021. 6

[6] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2, 3, 4, 6, 7

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 3

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 6

[10] Sungwon Hwang, Junha Hyung, and Jaegul Choo. Text2control3d: Controllable 3d avatar generation in neural radiance fields using geometry-guided text-to-image diffusion model, 2023. 4

[11] Junha Hyung, Jaeyo Shin, and Jaegul Choo. Magicapture: High-resolution multi-concept portrait customization, 2023. 4

[12] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 3

[13] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 6

[14] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[15] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022. 2

[16] Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models, 2023. 7

[17] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 4

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[19] 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*, pages 10684–10695, 2022. 3, 4

[20] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. 3, 4, 6

[21] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, 2022. 7

[22] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. 7

[23] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 3

[24] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. Pv3d: A 3d generative model for portrait video generation, 2023. 7

[25] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021. 7

[26] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[27] Hao Zhang, Yanbo Xu, Tianyuan Dai, Yu-Wing, and Tai Chi-Keung Tang. Fdnerf: Semantics-driven face reconstruction, prompt editing and relighting with diffusion models, 2023. 4

[28] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance, 2023. 4

[29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3, 4, 6

[30] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. pages 8568–8577, 2023. 4

[31] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset, 2022. 7