This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Controllable and Expressive One-Shot Video Head Swapping

Chaonan Ji Jinwei Qi Peng Zhang Bang Zhang Liefeng Bo
Alibaba Group

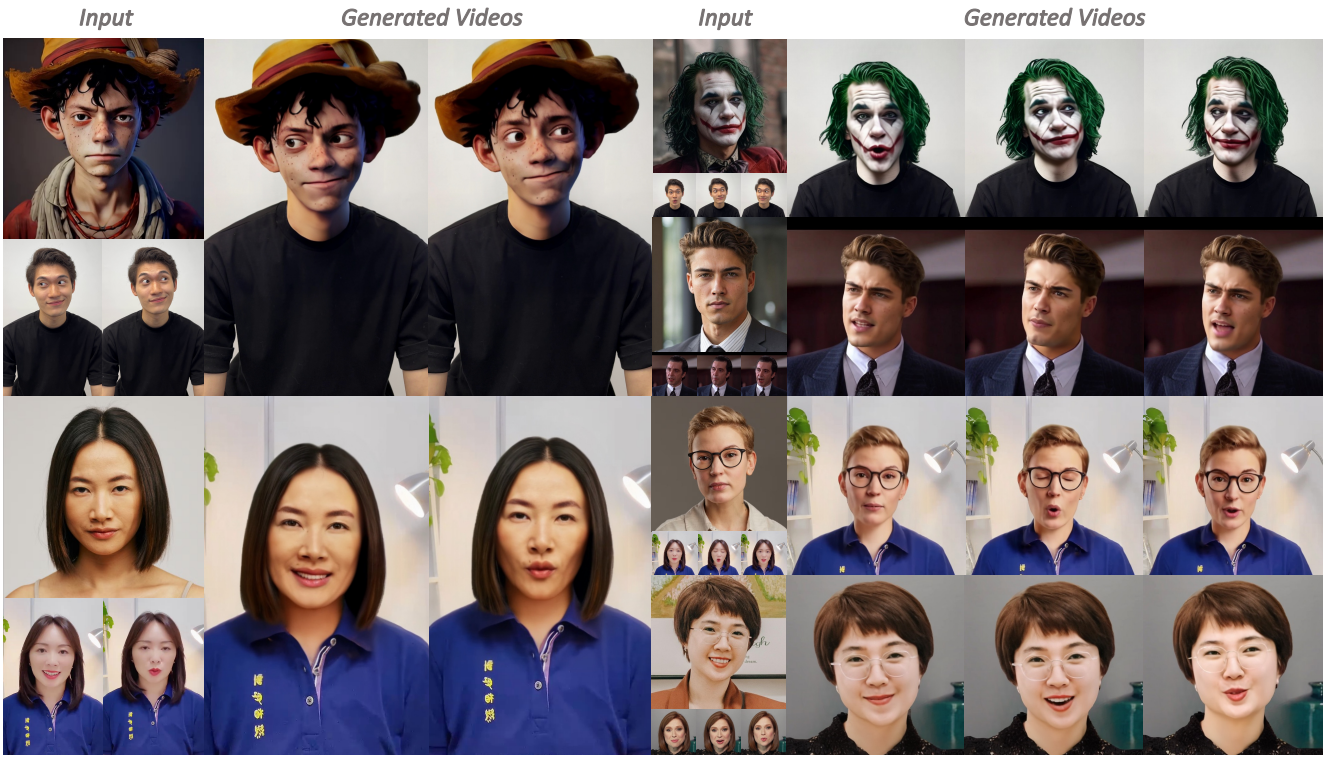

Figure 1. **Qualitative results of our method.** Given a reference image and video sequence as input, our model can generate high-fidelity head swapping results that accommodate diverse hairstyles, expressions, and identities.

Abstract

In this paper, we propose a novel diffusion-based multi-condition controllable framework for video head swapping, which seamlessly transplants a human head from a static image into a dynamic video, while preserving the original body and background of target video, and further allowing to tweak head expressions and movements during swapping as needed. Existing face-swapping methods mainly focus on localized facial replacement neglecting holistic head morphology, while head-swapping approaches struggling with hairstyle diversity and complex backgrounds, and none of these methods allow users to modify the transplanted head's expressions after swapping. To tackle these challenges, our method incorporates several innovative strategies through a unified latent diffusion paradigm. 1) Identity-preserving context fusion: We propose a shape-agnostic mask strategy to explicitly disentangle foreground head identity features from background/body contexts, combining hair enhancement strategy to achieve robust holistic head identity preservation across diverse hair types and complex backgrounds. 2) Expression-aware landmark retargeting and editing: We propose a disentangled 3DMM-driven retargeting module that decouples identity, expression, and head poses, minimizing the impact of original expressions in input images and supporting expression editing. While a scale-aware retargeting strategy is further employed to minimize cross-identity expression distortion for higher transfer precision. Experimental results demonstrate that our method excels in seamless background integration while preserving the identity of the source portrait, as well as showcasing superior expression transfer capabilities applicable to both real and

virtual characters.

1. Introduction

Video head swapping, in particular, seeks to seamlessly replace a source head in a target video while preserving the source individual's identity and target video's background and expression, offering immense application potential in film production, virtual reality and advertisement composition.

Face swapping has long been a topic of interest. Most face-swapping methods [5, 20, 22, 23, 34, 55, 63] aim at replacing localized facial regions by injecting identity features, yet they struggle to effectively transfer hairstyles, head shapes. In contrast, Video head swapping that focus on replacing the entire head in the target image, has been relatively less explored. StylePoseGAN [41] employs GAN inversion for head swapping but faces challenges in maintaining consistency of body and background; HeSer [42] utilizes two-stage training to decouple facial expressions and merge backgrounds, but it struggles with transferring hairstyles and inpainting complex backgrounds, often resulting in noticeable artifacts. Certain Diffusion-Based methodologies [40, 65] integrate identity features via self-attention mechanisms, which effectively facilitate the retention of identity characteristics. However, such methods are unable to generate images under specified background and body conditions.

Moreover, to ensure seamless head-swapping in video sequences, accurate and expressive facial expression transfer is indispensable. IPAdapter [65], DreamBooth [40] and HS-Diffusion [53] lacks the capability to control the facial expressions in the generated images. In contrast, AniPortrait [57] can transfer target expressions but is limited by the lack of expressiveness in facial templates. Follow-your-emoji [32] employs 3D landmarks for more expressive transfer but the results may be influenced by the reference image's original expressions. Expression transfer accuracy in both methods declines when there are significant differences in the proportional sizes of facial features between the source and target images. Neither method is adequate for head swapping.

In summary, current head swapping methods encounter two principal challenges: 1) preserving the identity of the source image (notably head shape and hairstyles) while achieving seamless integration with the target image's background; 2) accurately transferring expressions from the target video without being influenced by the facial proportions and expressions of the source image.

In this paper, we propose a controllable one-shot diffusion-based framework for video head swapping, that ensures high-quality background generation and identity consistency while allowing for expressive and controllable transferring of facial expressions and facilitating natural head movements. We reformulate head swapping as a conditional inpainting task for the head region, guided by identity features, background cues, and 3D facial landmarks. After segmenting and removing the head region of the target image, our model generates a new head to seamlessly integrate into the designated region while maintaining consistency with the surrounding background and body.

However, head segmentation may lead to shape leakage issues. To counter this, we introduce a shape-agnostic mask strategy designed to minimize head shape leakage issues by disrupting the segmented boundaries of the head. In addition, segmentation naturally creates a link between the hair and the body, particularly with long hair resting on the shoulders. We propose a hair enhancement strategy that disrupts the relationship between the hair and body, enabling the generated results to retain the hairstyle from the source image, supporting head swapping across diverse hairstyles.

Additionally, we introduce an expression-aware landmark retargeting module that decouples identity, expression, and pose, allowing for expression transfer and editing. This module eliminates expressions from the source image and adjusts driving landmarks according to the proportions of facial features. Specifically, we first employ a 3DMM [51] to fit the source image, obtaining a neutral expression template through expression coefficients adjustments. Subsequently, we determine the source landamrks by aligning the adjusted facial template's projected landmarks with 3D landmarks directly detected by the Mediapipe [29]. This process yields the source 3D landmarks in neutral expression, thus maintaining the expressive capacity of the Mediapipe landmarks while effectively removing the original expressions present in the source image. Building on this, we further adjust the expression intensity according to the size ratios of the facial features between the source and target individuals, which enhances the accuracy of the expression transfer and facilitating expression editing.

To conclude, our main contributions are the following:

- A novel controllable one-shot diffusion-based framework for video head swapping that ensures high-quality background generation and preserves identity while allowing for expressive facial expression transfer and editing.
- We propose a shape-agnostic mask strategy alongside a hair enhancement strategy that better preserves identity and supports diverse hairstyles in head swapping.
- We present an expression-aware landmark retargeting module that preserves expressive facial driving, removes the source image's expression influence, and accommodates expression transfer across portraits with different facial feature proportions.

2. Related Work

2.1. Face and Head Swapping

Face swapping [5, 20, 22, 23, 34, 55, 63, 72] has become a prominent research topic, aimed at fusing the identity information extracted from the source face to the target face. However, such methods typically do not adequately consider the head shape and hairstyle, which constrains the similarity of the swapped face to the source image. In contrast, head swapping [10, 25, 27, 41, 42, 53] that focuses on exchanging the entire head region, remains comparatively underexplored.

Deepfacelab [27] attempts to address the head swapping challenge by utilizing multiple reference images; however, the swapped results encounter artifacts in the fusion regions. GAN-based approaches [10, 25, 41] enable head swapping by editing localized face areas, yet often struggle with maintaining background consistency and exhibit difficulties in generalization. HeSer [42] integrates head driving with background fusion algorithms, achieving high-quality head swapping. However, it faces limitations in generating diverse hairstyles, transferring complex expressions and maintaining background. IPAdapter and related approaches [9, 40, 65] seek to incorporate head identity information and employ text prompts to generate diverse, identity-consistent images. However, these methods face challenges in precisely controlling backgrounds and expressions. HS-Diffusion [53] applies a semantic-mixing diffusion model guided by human parsing for head swapping, though it is unable to effectively transfer expressions. In response to these challenges, we propose a head swapping framework based on Latent Diffusion Models (LDMs) [38] that successfully integrates target background, diverse hairstyle, and expression transfer, facilitating high-quality video head swapping.

2.2. Expression transfer

GAN-based expression transfer has attracted substantial research interest over time. Previous methods [4, 11, 21, 36, 37, 44, 45, 49, 68] have primarily concentrated on lip-sync synthesis within the mouth region. To enhance the overall expressiveness of talking face, many methods [7, 8, 12, 13, 17, 19, 26, 28, 35, 46, 48, 52, 54, 56, 58, 62, 66, 69] attempt to encompass the entire head region.

Significant efforts have been dedicated to refining these control signals to enhance the precision of expression control. StyleAvatar [52] employs 3D Morphable Model (3DMM) [51] for facial expression transfer. Face vid2vid [54], SadTalker [69] along with LivePortrait [12] utilize implicit keypoints to achieve high-quality expression transfer. Furthermore, MegaPortrait [7], EMOPortrait [8], VISA [62], and LIA [56] adopts motion latent derived from images as control signals, which effectively bypass the constraints of 3DMM but introduce the risk of identity leakage. While these techniques display commendable expressiveness, they are inherently limited by the generative capacities of GANs, which often hinders their ability to faithfully transfer expressions.

Recently, Latent Diffusion Models (LDM) [38] have demonstrated superior generative capabilities, achieving high-quality video generation via text [3, 14, 15, 43, 73] or speech inputs [50]. However, the weak control signals pose challenges for precise expression control, thus many methods explore spatial-aligned control signals [1, 30–32, 57, 60, 61, 64, 70] by ControlNet [67] or PoseGuider [16]. MagicAnimate [64] employs densepose map as control signal, successfully transferring poses but struggling to accurately replicate expressions. AniPortrait [57] and MagicDance [1] utilizing facial landmarks derived from 3DMM, facilitating high-quality cross-identity image generation, and Follow-you-emoji [32] advances this methodology by incorporating 3D landmarks detected by Mediapipe [29], enabling more expressive expression transfers. Our approach builds upon these advancements by integrating 3DMM template into the 3D landmarks derived from Mediapipe [29], thereby achieving more precise expression transfer while maintaining identity consistency.

3. Method

3.1. Overview

In this section, we first present our framework in Sec.3.2. We then detail the Shape-Agnostic Mask and Hair Enhancement strategy in Sec.3.3. Lastly, we introduce the Expression-Aware Landmark Retargeting in Sec.3.4. The pipeline is illustrated in Fig.2.

3.2. Framework

Our framework is developed based on LDM [38] and is extended with background and expression conditions. The framework consists of the following modules: 1) **AppearanceNet:** It extracts identity information from the source image and injects it layer by layer into the DenoisingNet of the Stable Diffusion (SD) model through self-attention blocks, sharing the same architecture as DenoisingNet. 2) **PoseGuider:** It extracts multi-scale motion features and fuses them with the SD Unet features, establishing a spatial connection between the control signals and the generated image. 3) **MotionModule:** maintains temporal coherence and consistency across different frames. 4) **Image Encoder:** replaces the text encoder by encoding the reference image into a one-dimensional feature and injects it into the network through cross-attention blocks, providing additional global information from the reference image. The detail is shown in Fig.2

During training, given a reference image I_{ref} and a input

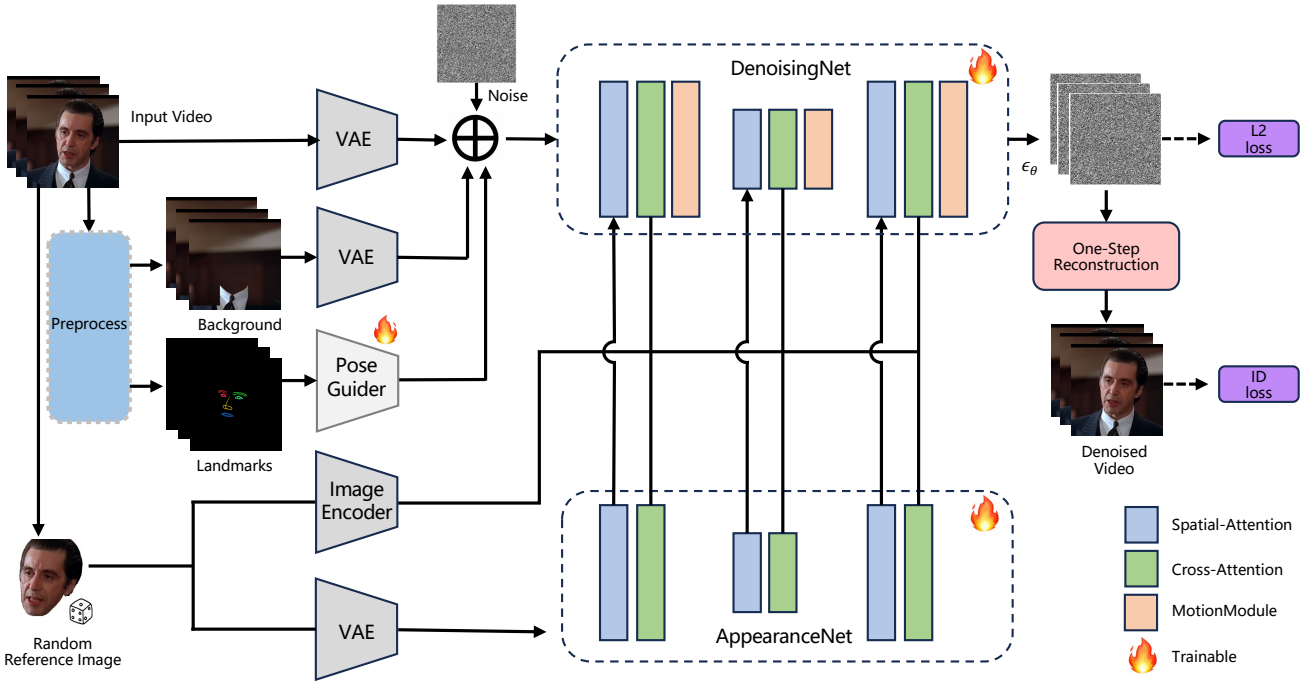

Figure 2. The framwork of our method. During the training stage, we begin by preprocessing the input video using the method outlined in Sec.3.3 to obtain the inpainted background and detect 3D landmarks uiltizing Mediapipe [29]. Both the inpainted background and the 3D landmarks serve as conditional inputs, and a frame is randomly selected as the reference image. We reconstruct the image from output noise and introducing additional ID losses in pixel level to enhance identity consistency.

video $I_d^{1:N}$ where N denotes the number of frames, we preprocess the input video to obtain the background sequence $I_b^{1:N}$ and extract driving landmarks $I_m^{1:N}$ to serve as the driving signal. For simplicity , we omit the superscript N. The training objective of SD can be written as:

$$\mathcal{L}_{LDM} = \mathop{\mathbb{E}}_{t,z_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t\right)\right\|^2\right] \tag{1}$$

where z_0 denotes the latent embedding of I_d, ϵ_θ is the predicted noise by SD and ϵ is the ground truth noise at timestep t. c is the condition embedding, which represents $c = (I_b, I_m, I_{ref})$ for our method.

To better preserve the identity, we recover the latent embeddings from the predicted noise ϵ_θ and decode it into an denoised video sequence $\hat{I}_d$ using VAE decoder. The detail of the ID loss will be explained in the supplementary materials. Subsequently, we compute the ID loss between the denoised video and the input video:

$$\begin{aligned}\mathcal{L}_{id} = \mathbb{E}\left[\left\|\mathcal{M}_{head} \odot (I_d - \hat{I}_d)\right\|^2\right] \\ + \mathbb{E}\left[1 - CosSim(\varepsilon_{id}(I_d), \varepsilon_{id}(\hat{I}_d))\right]\end{aligned} \tag{2}$$

where $\mathcal{M}_{head}$ is the head region mask and ε_{id} is a face recognition model [6] that encodes a face image into an identity embedding. Our total loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{LDM} + \mathcal{L}_{id} \tag{3}$$

During inference, given a reference image and a driving video, our method can generate high-quality head wrapping video with consistent background and expressive expression.

Condition Images Preprocessing To eliminate the influence of the background and body from the reference image, we use a human parsing algorithm [18] to retain only the reference image's head region as the input I_{ref} for the AppearanceNet. Additionally, we randomly scale and rotate the reference image to disrupt the positional relationship between the head and body, preventing identity leakage. In addition, we segment the foreground and clothing portions of the reference image, then employ an inpainting model [47] to inpaint the complete background. The inpainted background is recombined with the clothing portion to form a new background image I_b, which serves as the input to the network. The landmarks are detected by Mediapipe [29].

3.3. Shape-Agnostic Mask and Hair Enhancement

While the the aforementioned framework effectively produces head-swapping results, challenges remain in preserving identity of the reference image and generating diverse

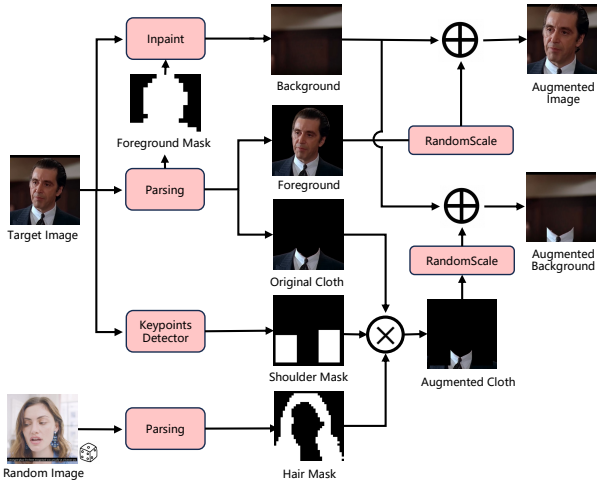

Figure 3. The detail of Shape-Agnostic Mask and Hair Enhancement Strategy. Given a target image, we first obtain the foreground and cloth masks. Then we apply a Shape-Agnostic Mask Strategy to disrupt the segmentation boundaries and obtain an inpainted background. The foreground is randomly scaled and combined with this background to create the augmented image used as training data. Additionally, a random long-haired image is selected to extract a hair mask, and a shoulder mask is created using shoulder keypoints detected by Mediapipe. The original cloth, shoulder mask, and hair mask are combined as the SD input.

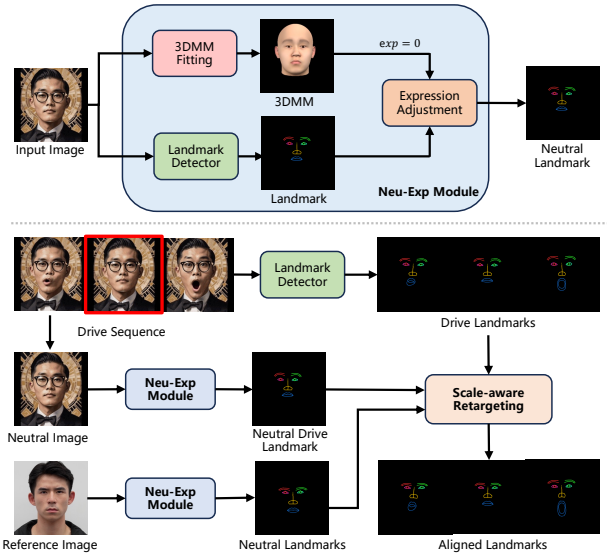

Figure 4. The detail of Expression-Aware Landmark Retargeting. The Neu-Exp module uses a 3DMM template to neutralize the expression in the input image, outputting neutral landmarks. For expression transfer, we select a nearly neutral frame from the drive sequence and process it with the Neu-Exp module to obtain Neutral drive landmarks. Then we use scale-aware retargeting to generate aligned landmarks.

hairstyles. These issues are attributed to identity leakage in the training data. Specially, due to the absence of authentic head-swapping videos, we utilize different frames from the same video as the reference image and the target image for training. Both the inputs I_{ref}, I_b of SD and target video I_d share the same identity. The inpainted backgrounds inadvertently encode head shape information through invisible segmentation boundaries (indistinguishable to the human eye but discernible by SD). In addition, this inherently introduce a bias that the generated head is constrained to remain within invisible boundaries, restricting hairstyle diversity.

To address these challenges, we propose a novel Shape-Agnostic Mask strategy. The details are shown in Fig.3. Given an input image I_d, we first obtain the corresponding foreground mask and then dilate it to obtain M_f. Subsequently, we define a new mask M_f^{new}, which shares the same dimensions as M_f and is composed of non-overlapping blocks of size $(k_h \times k_w)$. Each block in M_f^{new} is assigned a uniform value of 0 or 1, determined by the corresponding pixel values in M_f:

$$M_f^{new}(i,j) = \left\{ \begin{array}{ll} 1, if & any(M_f(i,j)) > 0 \\ 0, if & every(M_f(i,j)) = 0 \end{array} \right\} \quad (4)$$

where $M_f^{new}(i,j)$ represents value of the block at position (i,j). The $M_f^{new}(i,j)$ is used to generate the inpainted background. Since the edges are disrupted, the shape information of the reference image will not be leaked. In addition, during the training phase, we randomly scale the foreground to overcome the bias imposed by the inpainting model, allowing the generated head to extend beyond the limits of M_f^{new}. Specifically, we first preprocess the input video to obtain the foreground and inpainted background. The foreground is then randomly scaled and recomposed with the inpainted background to produce new training data. This strategy ensures that the foreground has a certain probability of covering areas beyond the original region, thus breaking through the invisible boundaries of the inpainted background. Consequently, during the inference phase, the SD can disregard the influence of invisible boundaries on hair shape and hairstyle.

Furthermore, the shoulder region often provides strong prior information about hairstyles. For example, in cases of long hair draping over shoulders, the cloth segmentation may create holes in the shoulder area reflecting the hair shape from the target image, introducing hairstyle priors during training. During testing, this leads to the generation of long hair in these gaps, even if the reference image depicts a short-haired individual. To address this issue, we propose a Hairstyle Enhancement Strategy, which involves 1) Collecting a set of long-hair videos, and randomly choosing a portrait image I_{hair} from this set during training. Then

we align its facial features with the target image, and obtain a hair region mask M_{hair}. 2) Using shoulder keypoints detected by Mediapipe, and generating random rectangular masks M_{rect} around these keypoints. 3) Eroding the original cloth region with M_{hair} and M_{rect} to disrupt hair prior. The final cloth mask can be represented as:

$$M_{cloth}^{new} = M_{cloth} \odot (1 - M_{hair}) \odot (1 - M_{rect}) \quad (5)$$

where M_{cloth} is original body mask and M_{cloth}^{new} is the augmented body mask. This approach effectively mitigates undesired hairstyle priors in the shoulder area.

3.4. Expression-Aware Landmark Retargeting

To achieve high-quality head-swapping results, effective and expressive expression transfer is essential. AniPortrait [57] employs a 3DMM to fit the reference image, efficiently decoupling identity, pose and expression, and reducing the reference image's expression influence. However, the expressiveness of the final result is constrained by the 3DMM's ability. In contrast, Follow-your-emoji [32] leverages 3D landmarks detected by Mediapipe [29], offering enhanced expressiveness. Nevertheless, its landmarks retargeting strategy does not address residual expressions in the reference image and struggles with cross-identity expression transfer when there are significant differences in facial feature proportions. To overcome these challenges, we propose an innovative Expression-Aware Landmark Retargeting strategy. The pipeline is shown in Fig.4. Given a reference image I_{ref} and a driving sequence I_d, the **Neu-Exp Module** first utilizes the Mediapipe detector to obtain reference landmarks L_{ref} and driving landmarks L_d, respectively. Then it fits the reference image using a 3DMM template [51] and projects the template to obtain the landmarks L_{ref}^{t}. Next, we set the expression coefficients of the template to 0 and reproject it to acquire the landmarks $L_{ref}^{t,exp=0}$. The neutral landmarks of reference image L_{ref}^{neu} can be represented as:

$$L_{ref}^{neu} = L_{ref} - L_{ref}^{t} + L_{ref}^{t,exp=0} \quad (6)$$

Subsequently, a frame depicting a neutral expression is selected from the driving sequence and undergoes the same processing to produce neutral landmarks L_d^{neu} of the driving sequence.

After obtaining neutral landmarks for both reference and driving images, we can further adjust the expression scale of various facial features through **Scale-aware Retargeting**, such as the eyes and mouth. Taking eyes as a case study, given the top and bottom landmark indices for the eyes, denoted as $i_{eye}^{top}, i_{eye}^{bot}$, we compute the expression magnitude adjustment ratio s_{eye} along yaw axis as follows:

$$s_{eye} = \frac{L_{ref}^{neu}[i_{eye}^{bot}, 1] - L_{ref}^{neu}[i_{eye}^{top}, 1]}{L_{d}^{neu}[i_{eye}^{bot}, 1] - L_{d}^{neu}[i_{eye}^{top}, 1]} \quad (7)$$

A similar procedure is applied to the mouth to obtain s_{mouth}. The scale-aware retargeting process can be formulated as:

$$\Delta L = L_d - L_d^{neu} \quad (8)$$

$$\begin{aligned} L = L_{ref}^{neu} + \Delta L + (s_{mouth} - 1) \odot \Delta L[:, i_{mouth}] \\ + (s_{eye} - 1) \odot \Delta L[:, i_{eye}] \end{aligned} \quad (9)$$

where L is the retargeted driving landmarks and i_{mouth}, i_{eye} is the landmarks indices of mouth and eyes, respectively. Benefiting from the disentanglement of identity, expression, and pose, as well as the control of expression magnitude, our method also enables expression editing. For more details, please refer to the supplementary material.

4. Experiment

4.1. Implementation Details

We train our model on the HDTF [71], Voxceleb [33], VFHQ [59] and TalkingHead1KH [24]. We first detect and crop the video's head region, then split the video into clips with durations of less than 30 seconds. To ensure quality, we utilize Image Quality Assessment (IQA) [2] to filter out clips with inferior image quality. Finally, our training dataset consists of 30K video clips from about 20K identities. During the training phase, we randomly sample two frames with different expressions from the video clips (one as the reference image and the other as the target image) and the images are resized as 512×512. We employ Mediapipe [29] detector to extract 3D landmarks from the target image and use lama as driving signals [47] to inpaint the background. The training details are explained in the supplementary materials.

4.2. Comparison with baselines

4.2.1. Qualitative results

We compare our method with previous state-of-the-art face swapping methods SimSwap [5], DiffSwap [72], head swapping method HeSer [42] and portrait animation methods Follow-your-emoji [32] and AniPortrait. HeSer utilizes third-party reproduced code as the official implementation is not available. HS-Diffusion is excluded as no public code is available. The results are presented in Fig.5. Face swapping methods [5, 72] often fail to effectively transfer head shapes and hairstyles, resulting in poor identity consistency. DiffSwap [72] generates blurry images. Head swapping method [42] transfers the entire head but may introduce background artifacts, inaccurate expression transfer and identity alteration. In contrast, our method maintains identity consistency and seamless foreground-background integration, accurately transferring target image expressions, even for uncommon portrait such as the

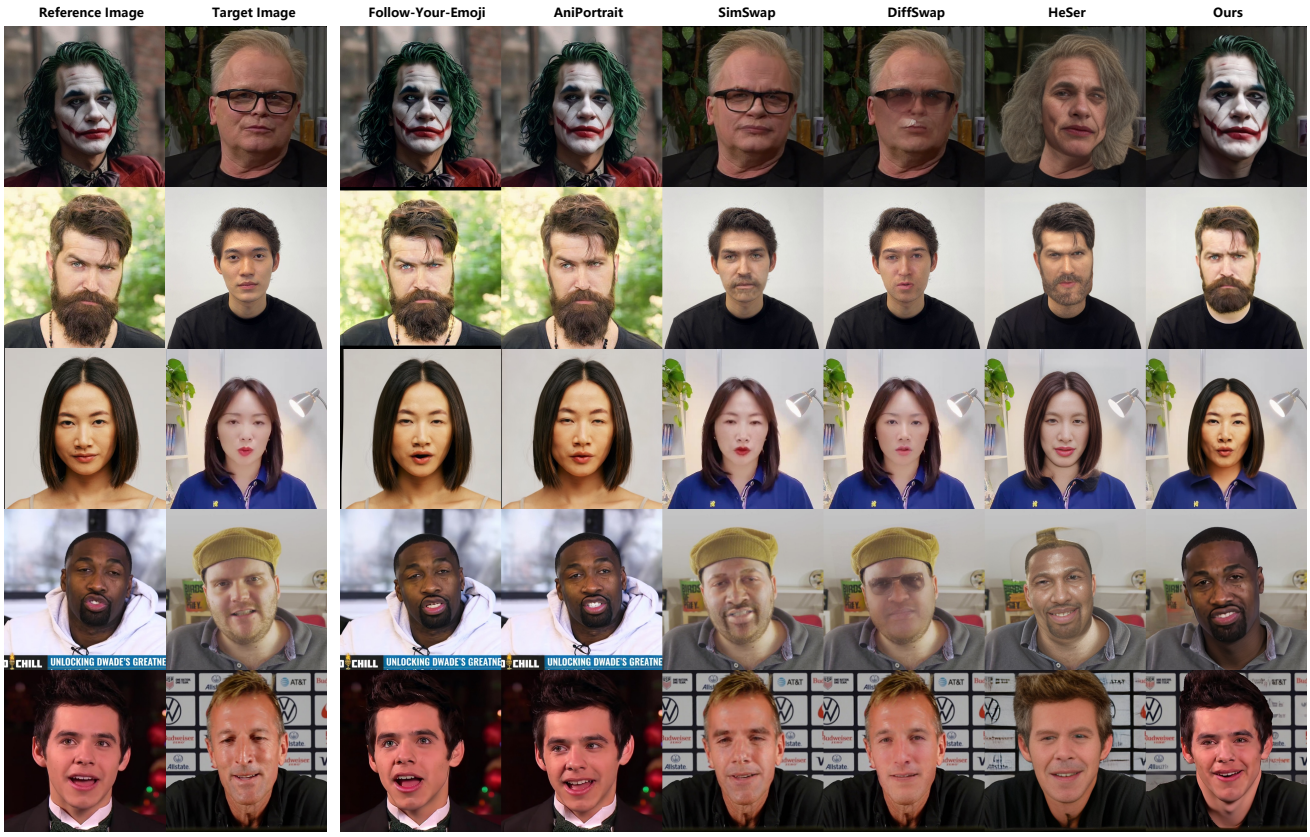

Figure 5. Qualitative comparion for head swapping.

Method	ID Similarity ↑	Pose ↓	Expression ↓
Aniportrait[1]	0.892	6.55	0.025
Emoji[1]	0.906	16.8	0.026
SimSwap[2]	0.592	0.702	0.005
DiffSwap[2]	0.397	1.144	0.008
HeSer[3]	0.319	6.840	0.022
Ours	0.895	9.83	0.014

Table 1. The quantitative results. [1] denotes portrait animation methods, [2] denotes face swapping methods, and [3] indicates head swapping methods. The optimal results are marked in red, the second-best results are marked in blue. Emoji denotes Follow-Your-Emoji.

Joker at the first row. For comparison results of portrait animation , please refer to the supplementary materials.

4.2.2. Quantitative results

For quantitative evaluation, we randomly selecte 200 videos with unique identities from the VFHQ [59] dataset, extracting 48 frames from each video at intervals of 2 as target videos. Additionally, we chose 200 portrait images from VFHQ as reference images, ensuring different identities from the choosed videos. We evaluate ID Similarity, pose error, expression error, and FID. ID similarity is measured by extracting indetity features with ArcFace [6] and calculating cosine similarity from both generated and reference images. Pose error is assessed using a pretrained head pose estimator [39] to extract head poses from both generated results and target videos, followed by L2 distance calculation. Expression error is evaluated by extracting expression coefficients with a facial blendshape extractor [29] from both sets and computing the L2 distance.

The results are shown in Tab.1. For expression, our method achieves smaller errors compared to analogous method Follow-your-Emoji [32] and AniPortrait [57]. For pose error, our method is comparable to HeSer and AniPortrait, with differences likely attributable to the variations in the 3D landmarks employed. Our approach outperforms Follow-Your-Emoji that also utilizes Mediapipe landmarks. Face-swapping methods obatin smallest errors as they tend to make subtle modifications to the target image, whereas our approach involves fully regenerating the head. For ID similarity, our method substantially surpasses traditional face-swapping and head-swapping approaches by comprehensively preserving head features. While head-driven methods inherently exhibit superior ID retention without the

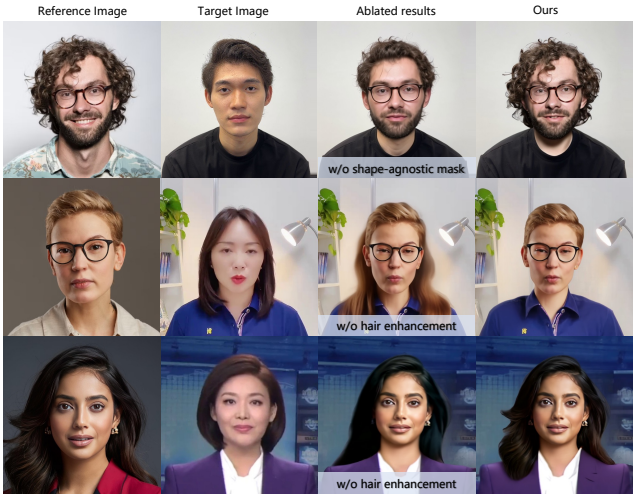

Figure 6. The ablated results of Shape-Agnostic Mask and Hair Enhancement Strategy.

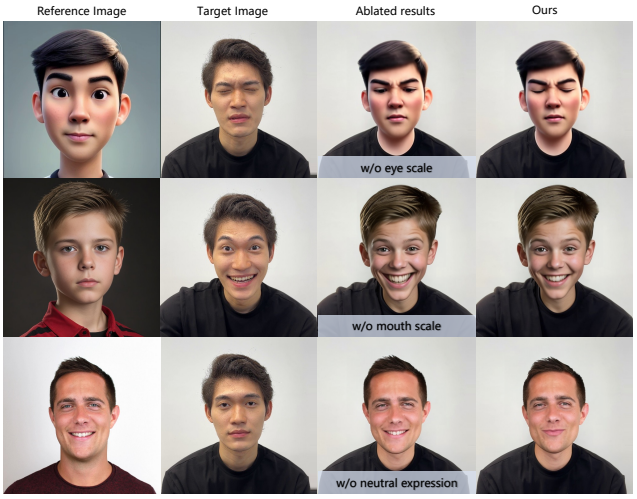

Figure 7. The ablated results of Expression-Aware Landmark Retargeting.

need to integrate new background and body integration, our approach achieves comparable accuracy under these conditions. This highlights our method's robust capability for identity preservation in head-swapping scenarios.

4.3. Ablation study

In this section, we analyze the effectiveness of the Shape-Agnostic Mask and Hair Enhancement strategy, as well as the Expression-Aware Landmark Retargeting module.

Effectiveness of the Shape-Agnostic Mask and Hair Enhancement. To assess the effectiveness of the two strategies, we conduct experiments by disabling each one separately during training. The comparative results are illustrated in Fig.6. When Shape-Agnostic Mask strategy is omitted, the generated head region is disproportionately influenced by the head shape of the original driving video, leading to head shape mismatch with the reference image. Moreover, the generated head is strictly confined within the head boundaries of the driving video, restricting the region available for hair generation. Without Hair Enhancement strategy, the model generates undesirable hair artifacts due to incorrect shoulder-related priors as shown in the second row. It also fails to naturally generate hair in front of the shoulders without shoulder mask, as shown in the third row. By contrast, our model maintains better identity consistency and is capable of generating diverse hairstyles.

	Expression ↓
ours(w/o Neu-Exp Module)	0.025
ours(w/o Scale-aware Retargeting)	0.015
ours	**0.014**

Table 2. The quantitative results of the Expression-Aware Landmark Retargeting.

Effectiveness of Expression-Aware Landmark Retargeting. To verify the effectiveness of the Expression-Aware Landmark Retargeting module, we conduct experiments by discarding Neu-Exp Module and Scale-aware Retargeting repectively during inference. The Qualitative results are presented in Fig.7. In the absence of Neu-Exp Module, the driven expressions overlay on the existing expressions of the reference image (for example, a smile), which reduces the accuracy of expression transfer. Furthermore, without Scale-aware Retargeting, the accuracy of expression transfer diminishes, especially between subjects with significant differences in facial feature sizes; For instance, Mismatched eye sizes may result in improper eye closure, as shown in the first row. Variations in mouth size between the reference image and driving video can lead to exaggerated facial expressions as shown in the second row.The quantitative comparison is detailed in Tab.2

5. Conclusion

we present a novel diffusion-based multi-condition controllable framework for video head swapping that ensures background and identity consistency while enabling expressive and flexible facial expression transfer and editing. By integrating innovative strategies such as shape-agnostic masking and a novel hair enhancement strategy, our method reduces identity leakage and supports the generation of diverse hairstyles. The expression-aware landmark retargeting module enhances the accuracy and expressiveness of cross-identity expression transfer, and supports expression transfer. Our framework achieves seamless head swapping across various hairstyles and complex backgrounds, and precise expression transfer. Experimental results confirm the method's superiority on a variety of scenes.

References

[1] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 3

[2] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. 6

[3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *CoRR*, abs/2310.19512, 2023. 3

[4] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 538–553. Springer, 2018. 3

[5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2003–2011. ACM, 2020. 2, 3, 6

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. 4, 7

[7] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 2663–2671. ACM, 2022. 3

[8] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8498–8507. IEEE, 2024. 3

[9] Zhongyi Fan, Zixin Yin, Gang Li, Yibing Zhan, and Heliang Zheng. Dreambooth++: Boosting subject-driven generation via region-level references packing. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 11013–11021. ACM, 2024. 3

[10] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J. Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7713–7722. IEEE, 2022. 3

[11] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, and Jingdong Wang. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 1505–1515. IEEE, 2023. 3

[12] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *CoRR*, abs/2407.03168, 2024. 3

[13] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5764–5774. IEEE, 2021. 3

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[15] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. 3

[16] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8153–8163. IEEE, 2024. 3

[17] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. EAMM: one-shot emotional talking face via audio-based emotion-aware motion model. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 61:1–61:10. ACM, 2022. 3

[18] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 4

[19] Taekyung Ki, Dongchan Min, and Gyeongsu Chae. FLOAT: generative motion latent flow matching for audio-driven talking portrait. *CoRR*, abs/2412.01064, 2024. 3

[20] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. Smooth-swap: A simple enhancement for face-swapping with smoothness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10769–10778. IEEE, 2022. 2, 3

[21] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *CoRR*, abs/1801.01442, 2018. 3

[22] Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, and Ran He. Faceinpainter: High fidelity face adaptation to heterogeneous domains. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5089–5098. Computer Vision Foundation / IEEE, 2021. 2, 3

[23] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *CoRR*, abs/1912.13457, 2019. 2, 3

[24] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: mask-aware transformer for large hole image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10748–10758. IEEE, 2022. 6

[25] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: mask-aware transformer for large hole image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10748–10758. IEEE, 2022. 3

[26] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3377–3386. IEEE, 2022. 3

[27] Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognit.*, 141:109628, 2023. 3

[28] 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. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 6696–6705. ACM, 2024. 3

[29] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. 2, 3, 4, 6, 7

[30] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 4117–4125. AAAI Press, 2024. 3

[31] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-your-click: Open-domain regional image animation via short prompts. *CoRR*, abs/2403.08268, 2024.

[32] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, pages 110:1–110:12. ACM, 2024. 2, 3, 6, 7

[33] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017*, pages 2616–2620. ISCA, 2017. 6

[34] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsganv2: Improved subject agnostic face swapping and reenactment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):560–575, 2023. 2, 3

[35] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-Ming Yan. DPE: disentanglement of pose and expression for general video portrait editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 427–436. IEEE, 2023. 3

[36] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 2062–2070. AAAI Press, 2022. 3

[37] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 3

[39] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2074–2083. Computer Vision Foundation / IEEE Computer Society, 2018. 7

[40] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22500–22510. IEEE, 2023. 2, 3

[41] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image syn-

thesis of humans from a single monocular view. *CoRR*, abs/2102.11263, 2021. 2, 3

[42] Changyong Shu, Hemao Wu, Hang Zhou, Jiaming Liu, Zhibin Hong, Changxing Ding, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Few-shot head swapping in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10779–10788. IEEE, 2022. 2, 3, 6

[43] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[44] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Trans. Inf. Forensics Secur.*, 17:585–598, 2022. 3

[45] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 919–925. ijcai.org, 2019. 3

[46] Yasheng Sun, Hang Zhou, Ziwei Liu, and Hideki Koike. Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1018–1024. ijcai.org, 2021. 3

[47] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3172–3182. IEEE, 2022. 4, 6

[48] Shuai Tan, Bin Ji, and Ye Pan. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26307–26317. IEEE, 2024. 3

[49] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, pages 716–731. Springer, 2020. 3

[50] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIII*, pages 244–260. Springer, 2024. 3

[51] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20301–20310. IEEE, 2022. 2, 3, 6

[52] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pages 67:1–67:10. ACM, 2023. 3

[53] Qinghe Wang, Lijie Liu, Miao Hua, Pengfei Zhu, Wangmeng Zuo, Qinghua Hu, Huchuan Lu, and Bing Cao. Hs-diffusion: Semantic-mixing diffusion for head swapping, 2023. 2, 3

[54] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10039–10049. Computer Vision Foundation / IEEE, 2021. 3

[55] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1136–1142. ijcai.org, 2021. 2, 3

[56] Yaohui Wang, Di Yang, François Brémond, and Antitza Dantcheva. LIA: latent image animator. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10829–10844, 2024. 3

[57] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *CoRR*, abs/2403.17694, 2024. 2, 3, 6, 7

[58] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1478–1486. ACM, 2021. 3

[59] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 656–665. IEEE, 2022. 6, 7

[60] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, page 115. ACM, 2024. 3

[61] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Trans. Vis. Comput. Graph.*, 31 (2):1526–1541, 2025. 3

[62] 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. In *Advances in Neural Information*

Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. 3

[63] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7632–7641. IEEE, 2022. 2, 3

[64] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 1481–1490. IEEE, 2024. 3

[65] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721, 2023. 2, 3

[66] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22096–22105. IEEE, 2023. 3

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3813–3824. IEEE, 2023. 3

[68] Longhao Zhang, Shuang Liang, Zhipeng Ge, and Tianshu Hu. Personatalk: Bring attention to your persona in visual dubbing. In *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, pages 108:1–108:9. ACM, 2024. 3

[69] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8652–8661. IEEE, 2023. 3

[70] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[71] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3661–3670. Computer Vision Foundation / IEEE, 2021. 6

[72] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8568–8577. IEEE, 2023. 3, 6

[73] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *CoRR*, abs/2211.11018, 2022. 3