FACESHOT: BRING ANY CHARACTER INTO LIFE

Anonymous authors
Paper under double-blind review

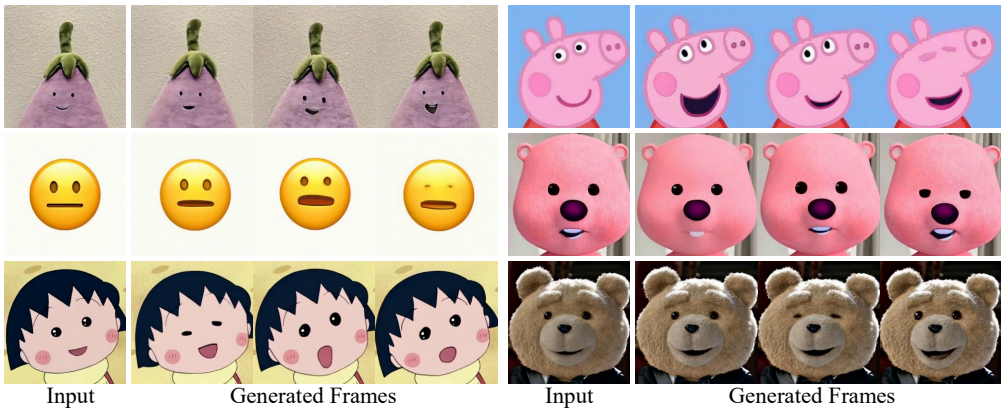

Figure 1: Visualization results of our **FaceShot**. Given any character and any driven video, FaceShot effectively captures subtle facial expressions and generates stable animations for each character. Especially for non-human characters, such as emojis and toys, FaceShot demonstrates remarkable animation capabilities.

ABSTRACT

Portrait animation generates dynamic, realistic videos by mimicking facial expressions from a driven video. However, existing landmark-based methods are constrained by facial landmark detection and motion transfer limitations, resulting in suboptimal performance. In this paper, we present *FaceShot*, a novel training-free framework designed to animate any character from any driven video, human or non-human, with unprecedented robustness and stability. We achieve this by offering precise and robust landmark results from an appearance guided landmark matching module and a relative motion transfer module. Together, these components harness the robust semantic correspondences of latent diffusion models to deliver landmarks across a wide range of character types, all without requiring fine-tuning or retraining. With this powerful generalization capability, FaceShot can significantly extend the application of portrait animation by breaking the limitation of landmark detection for any character and driven video. Furthermore, FaceShot is compatible with any landmark-driven animation model, enhancing the realism and consistency of animations while significantly improving overall performance. Extensive experiments on our newly constructed character benchmark CABench confirm that FaceShot consistently surpasses state-of-the-art approaches across any character domain, setting a new standard for open-domain portrait animation. Our code will be publicly available. More results are available at our anonymous project website `https://faceshot2024.github.io/faceshot/`.

1 INTRODUCTION

Facial motion capture technology is widely used to animate 2D and 3D characters in film production, game development, and VTubing. However, it often requires specialized equipment and significant manual effort for character modeling and rigging. Recent advancements in generative models, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) and diffusion models (Rombach et al., 2022), have demonstrated impressive results in portrait animation (Guo et al., 2024; Xie et al., 2024; Ma et al., 2024). These methods depend on facial landmark recognition and can generalize to certain portraits whose facial landmarks can be well detected. However, many non-human characters

created by designers, such as emojis, animals and toys, often exhibit significantly different facial features compared to human portraits, resulting in landmark recognition failures and negatively impacting the animation process.

Naturally, the performance of landmark-driven methods is constrained by the generalization capability of facial landmark detection (Zhou et al., 2023; Yang et al., 2023) and landmark motion transfer models (Booth et al., 2016). Due to the supervised training paradigm and limited datasets, these landmark-driven models struggle to adapt to non-human characters, leading to suboptimal animation performance. As shown in Figure 2 (line 1), due to inaccurate landmark results, the model even generates human mouth and facial features on a dog.

In this paper, we propose **FaceShot**, a novel portrait animation framework capable of animating any character from any driven video without the need for training. Our visualizations (Figure 1) demonstrate that FaceShot produces vivid and stable animations for various characters, particularly for non-human characters. This is achieved through three key components: (1) the appearance guided landmark matching module, (2) the relative landmark motion transfer module, and (3) the character animation model. For the first component, we inject appearance priors into diffusion features and leverage their strong semantic correspondences to match the landmarks. For the second component, we introduce a theoretical algorithm to capture subtle expressions between frames and generate the landmark sequence from the driven video. Finally, for the third component, we input the landmark sequence of the target character into a pre-trained model to generate the animation. As shown in Figure 2 (line 2), FaceShot provides the reasonable animation result by offering the precise landmarks of non-human character. Furthermore, FaceShot is compatible with any landmark-driven model as a plugin, improving their performance on non-human characters.

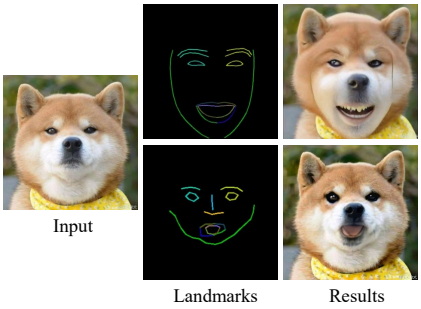

Figure 2: Animation results of inaccurate (line 1) and accurate (line 2) landmarks.

Moreover, to address the absence of a benchmark for character animation, we establish a new benchmark that contains 46 characters and 24 human head videos. Qualitative and quantitative evaluations on CABench demonstrate that FaceShot excels in animating characters, especially in non-human characters, outperforming existing portrait animation methods. Additionally, ablation studies validate the effectiveness and superiority of our framework, providing valuable insights for the community. Furthermore, we provide the animation results of FaceShot from non-human driven videos, bringing a potential solution to the community for open-domain portrait animation.

The main contributions of this paper are as follows:

- We propose FaceShot, a novel portrait animation framework capable of animating any character from any driven video.
- FaceShot generates precise landmark results for any character without the need for training and can be seamlessly integrated as a plugin with any landmark-driven animation model.
- FaceShot breaks the limitations of landmark detection and motion transfer, bringing a potential solution to the community for open-domain portrait animation.
- We establish CABench, a benchmark with diverse characters for comprehensive evaluation. Experiments show that FaceShot outperforms state-of-the-art (SOTA) approaches.

2 Related Work

2.1 Portrait Animation

Portrait animation has been extensively explored, with early methods primarily relying on GANs (Goodfellow et al., 2020) for motion generation through warping and rendering (Drobyshev et al., 2022; Siarohin et al., 2019). These approaches used pose predictors to guide facial warping, but struggled with stability and generating detailed facial features across diverse domains (Hong et al., 2022; Wang et al., 2021; Zhao & Zhang, 2022). Recently, latent diffusion models (LDMs) (Rombach

et al., 2022; Ramesh et al., 2022; Shen et al., 2023) have improved images quality and efficiency by operating in latent space. They (Niu et al., 2024; Wei et al., 2024) have been successfully applied to portrait animation tasks, producing photorealistic and temporally consistent results. Approaches like AniPortrait (Wei et al., 2024) and MegActor (Yang et al., 2024) leverage diffusion models to animate facial landmarks, while other methods, such as MOFA-Video (Niu et al., 2024), V-Express (Wang et al., 2024a), and Follow-Your-Emoji (Ma et al., 2024), enhance motion control with expression-aware landmarks. However, most of these methods still rely on landmark recognition, limiting their generalization to non-human characters. In our paper, we focus on animating any character from any driven video using precise landmark results.

2.2 Facial Landmark Detection

Facial Landmark detection aims to detect key points in given face. Traditional methods (Cootes et al., 2000; 2001; Dollár et al., 2010; Kowalski et al., 2017) often construct a shape model for each key point and perform iterative searches to match the landmarks. With the development of deep networks, some methods (Sun et al., 2013; Zhou et al., 2013; Wu et al., 2018; 2017) select a series of coarse to fine cascaded networks to perform direct regression on the landmarks. Another trend, Huang et al. (2020); Zhou et al. (2023); Merget et al. (2018); Kumar et al. (2020) predict the heatmap of each point for indirect regression, improving the accuracy of landmark detection. Recently, Yang et al. (2023); Xu et al. (2022); Li et al. (2024) collect larger datasets and train bigger models for open-domain landmark detection. However, due to the supervised training paradigm and limited dataset, these methods are difficult to perfectly detect the landmark of non-human characters. In our paper, we turn to a training-free landmark detection pipeline through the strong semantic correspondence in diffusion features (Tang et al., 2023; Hedlin et al., 2024; Luo et al., 2024), aiming to provide robust and precise landmark results for non-human characters.

2.3 Image to Video Generation

Image to video generation has gained significant attention in recent years due to its potential in various applications, such as image animation (Dai et al., 2023; Gong et al., 2024; Ni et al., 2023; Guo et al., 2023) and video synthesis (Blattmann et al., 2023b; Wang et al., 2024b; Ruan et al., 2023). Since diffusion models demonstrate the powerful image generation capabilities, Zhang et al. (2024); Shi et al. (2024); Xing et al. (2023); Ma et al. (2024) achieve image animation by inserting temporal layers into a pre-trained 2D UNet and fine-tuning it with video data. Furthermore, some methods (Zhang et al., 2023a; Blattmann et al., 2023a) have constructed their own I2V models and performed full training with large amounts of high-quality data, demonstrating strong competitiveness. In our work, we utilize stable video diffusion (Blattmann et al., 2023a) as our base animation model.

3 Method

The framework of FaceShot is depicted in Figure 3. We first introduce the foundational concepts of diffusion models in Section 3.1. The framework comprises three key components: appearance guided landmark matching, relative landmark motion transfer, and character animation generation, which are explained in detail in Section 3.2.

3.1 Preliminary

In FaceShot, we utilize Latent Diffusion Models (LDMs) (Rombach et al., 2022) for landmark matching, which consist of a Variational Auto-Encoder (VAE) (Kingma, 2013), a CLIP text encoder (Radford et al., 2021), and a denoising U-Net (Ronneberger et al., 2015). Compared to pixel-based diffusion models, LDMs use the VAE encoder $\mathcal{E}$ to encode the input image $\mathbf{x}$ into a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$. The VAE decoder $\mathcal{D}$ then reconstructs the image by decoding the latent representation, $\mathbf{x} = \mathcal{D}(\mathbf{z})$.

To train the denoising U-Net ϵ_θ, the objective typically minimizes the Mean Square Error (MSE) loss $\mathcal{L}$ at each time step t, as follows:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{c}, t} \|\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) - \epsilon_t\|^2, \tag{1}$$

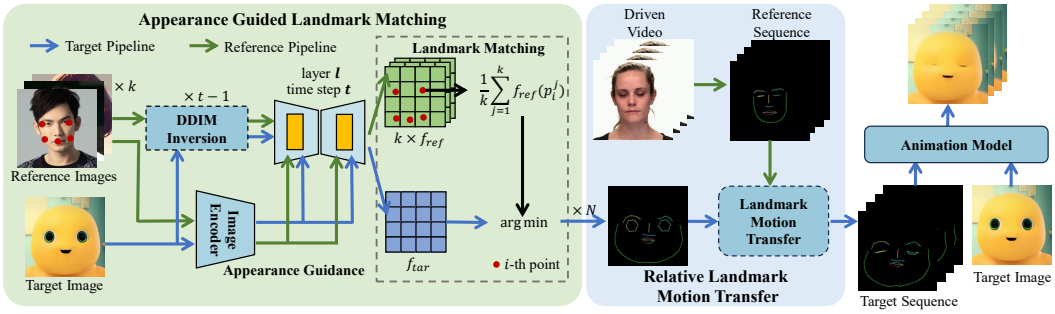

Figure 3: The framework of FaceShot first generates precise facial landmarks of the target character with appearance guidance. Next, a relative landmark motion transfer module is applied to generate the landmark sequence. Finally, this landmark sequence is input into an animation model to animate any character from any driving video.

where $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$ is the noisy latent at time step t. Here, ϵ_t represents the added Gaussian noise, and $\mathbf{c}$ is the text condition, typically processed by the U-Net's cross-attention module.

The Denoising Diffusion Implicit Model (DDIM) enables the inversion of the latent variable z_0 to z_t in a deterministic manner. The formula is as follows:

$$z_t = \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} z_{t-1} + \left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right) \cdot \epsilon_\theta\left(z_{t-1}, t-1, c, c_i\right), \tag{2}$$

where c_i represents the image prompt. In our approach, we utilize the latent space features at t-th time step and l-th layer of U-Net from DDIM inversion for landmark matching.

3.2 FaceShot: Bring any Character into Life

Facial landmarks play a critical role in determining the performance of portrait animation. Prior methods (Yang et al., 2023; Xu et al., 2022; Zhou et al., 2023) have either curated more diverse public datasets or introduced new loss functions during training to improve the generalization of landmark detection. However, within a supervised training paradigm, these detectors struggle to generalize to non-human characters outside the distribution of the training dataset. To address this, we first propose an appearance guided landmark matching module to generate the precise landmarks. Second, we provide the relative landmark motion transfer module to obtain the target landmark sequence from a driven video. Finally, a character animation model is employed as our base model to animate target characters.

Appearance Guided Landmark Matching. Tang et al. (2023); Hedlin et al. (2024); Luo et al. (2024) demonstrate the strong semantic correspondence between diffusion features, where simple feature matching can map the target point p' on the target image I_{tar} to a similar point p on the reference image I_{ref}. However, appearance discrepancies across different domains often result in mismatches, as shown in Figure 4 (a), where the points on the left eye and right ear are incorrectly matched. A natural solution is to inject prior appearance knowledge through inference using a domain-specific diffusion model. As shown in Figure 4 (b), the points are correctly matched.

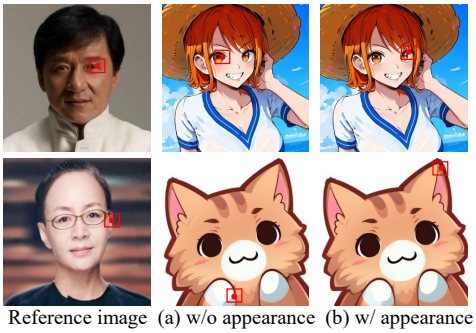

Figure 4: Visualizations of point matching with (w/) or without (w/o) appearance guidance by a domain-specific diffusion model. We highlight it with a red box for better visibility.

Since tuning a diffusion model for each target image is costly, inspired by IP-Adapter (Ye et al., 2023), we utilize image prompts to provide appearance guidance. Specifically, we treat the reference image I_{ref} and the target image I_{tar} as image prompts, denoted as c_{ref} and c_{tar}, respectively. We apply the DDIM inversion process to obtain deterministic diffusion features f_{ref} and f_{tar} from I_{ref} and

I_{tar} at time step t and the l-th layer of the U-Net:

$$\begin{aligned} f_{tar} &= F_l(\epsilon_\theta(z_{tar}^{t-1}, t-1, c, c_{ref})), \\ f_{ref} &= F_l(\epsilon_\theta(z_{ref}^{t-1}, t-1, c, c_{tar})), \end{aligned} \tag{3}$$

where z_{tar}^{t-1} and z_{ref}^{t-1} are iteratively sampled by Eq. 2 from $z_{tar}^0 = \mathcal{E}(I_{tar})$ and $z_{ref}^0 = \mathcal{E}(I_{ref})$ using the text prompt c and the image prompts c_{ref} and c_{tar}, respectively. F_l denotes the function that extracts the output feature at the l-th layer of the U-Net.

After obtaining the diffusion features $f_{ref} \in \mathbb{R}^{1\times C_l \times h_l \times w_l}$ and $f_{tar} \in \mathbb{R}^{1\times C_l \times h_l \times w_l}$, we upsample them to $f'_{ref} \in \mathbb{R}^{1\times C_l \times H \times W}$ and $f'_{tar} \in \mathbb{R}^{1\times C_l \times H \times W}$ to match the resolution of I_{ref} and I_{tar}. To improve performance and stability, we construct the invariant feature of the i-th landmark point p_i from k reference images to match the corresponding point p'_i in the target image, as follows:

$$p'_i = \arg\min_{p_{tar}} d_{cos}\left(\frac{1}{k}\sum_{j=1}^{k} f_{ref}^j(p_i^j), f_{tar}(p_{tar})\right), \tag{4}$$

where $f(p) \in \mathbb{R}^{1\times C_l}$ represents the diffusion feature vector at point p, and d_{cos} denotes the cosine distance. Here, p_{tar} refers to points in the target feature map. Finally, we denote the matched landmark points of the target image as $L_{tar}^0 = \{p'_i \mid i = 1, \ldots, N\}$, where N represents the number of facial landmarks.

Appearance Gallery. Moreover, we introduce an appearance gallery $G = [G_e, G_m, G_n, G_{eb}, G_{fb}]$, which is a collection of five prior components—*eyes, mouth, nose, eyebrows*, and *face boundary*—across various domains, with each domain containing k images. For a target image I_{tar}, we reconstruct the reference image as $I_{ref} = [G_e^*, G_m^*, G_n^*, G_{eb}^*, G_{fb}^*]$ by matching I_{tar} with the closest domain in the appearance gallery G, thereby explicitly reducing the appearance discrepancy between the reference and target images, as shown in Figure 5.

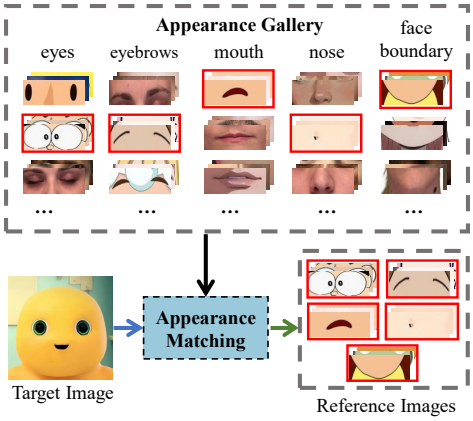

Figure 5: Illustration of our appearance gallery. We output the closest domains for each target image to reduce the appearance discrepancy.

Relative Landmark Motion Transfer. Currently, Niu et al. (2024); Wei et al. (2024); Ma et al. (2024) utilize 3D Morphable Models (3DMM) (Booth et al., 2016) to generate the landmark sequence of the target image by applying 3D face parameters. However, 3DMM-based methods often struggle to generalize to non-human character faces due to the limited number of high-quality 3D data and their inability to capture subtle expression movements (Retsinas et al., 2024). As shown in Figure 6, the head shapes of the 3D face are not well aligned with the input images, and subtle movements, such as eye closures, are absent in the i-th frame. To address these limitations, we propose a relative landmark motion transfer method, designed to generate a landmark sequence L_{tar} from L_{tar}^0 and the reference sequence L_{ref}.

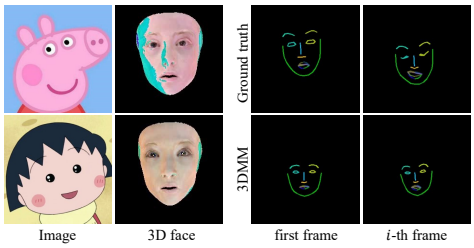

Figure 6: Visualizations of 3D face and motion transfer results using 3DMM.

Our theoretical method consists of two stages, which transfer the global and local motion of facial landmarks, respectively. In the first stage, global motion is defined as the translation and rotation of the global rectangular coordinate systems for L_{ref}^m and L_{ref}^0 in the m-th and 0-th frames. The global rectangular coordinate system is constructed by the origin O and angle θ, which are calculated from the endpoints of the face boundary. The global parameters of L_{tar}^m can be obtained from L_{tar}^0 as follows:

$$\begin{aligned} O_{tar}^m &= O_{tar}^0 + (O_{ref}^m - O_{ref}^0), \\ \theta_{tar}^m &= \theta_{tar}^0 + (\theta_{ref}^m - \theta_{ref}^0). \end{aligned} \tag{5}$$

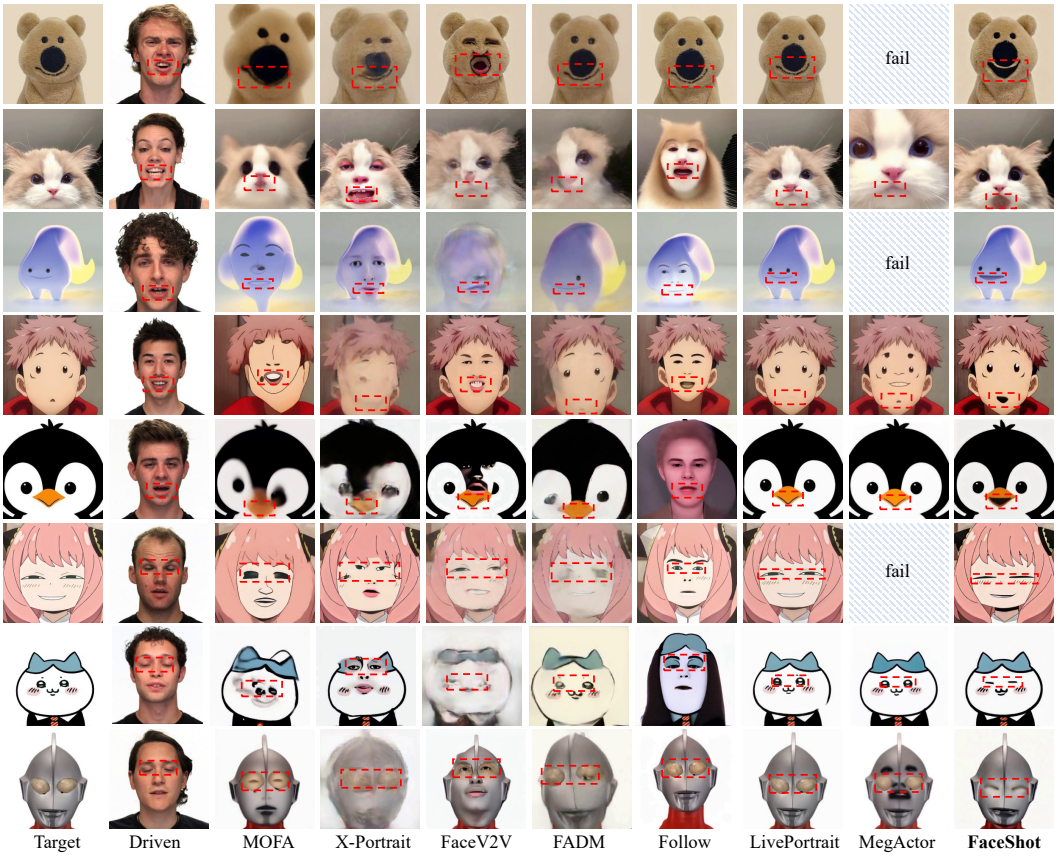

Figure 7: Qualitative comparison with SOTA portrait animation methods. Slash boxes represent that the method has fail to generate animation for this character.

In stage two, we first apply the algorithm from stage one to five local facial parts, including the eyes, mouth, nose, eyebrows, and face boundary, to transfer relative motion across the entire face, such as raising eyebrows. Additionally, we constrain the relative motion within a reasonable range using the scale factor b_{tar}/b_{ref}, where b represents the distance from the origin to the boundary of each part. Next, we model the movement of each point as follows:

$$p'_{m,i} = \left(\frac{p_{m,i}[0]}{p_{0,i}[0]} \cdot p'_{0,i}[0], \frac{p_{m,i}[1]}{p_{0,i}[1]} \cdot p'_{0,i}[1] \right), \tag{6}$$

where $p'_{m,i}$ and $p_{m,i}$ represent the coordinates of the m-th frame and i-th point of the target and reference landmarks within each local part in the corresponding local rectangular coordinate system. This simple yet effective theoretical design enables us to obtain the landmark sequence $L_{tar} = \{L^j_{tar} \mid j = 1, \ldots, M\}$ for any characters from a reference sequence, where M represents the number of video frames. Additionally, since extracting landmark sequences from non-human driven videos is difficult and unstable for landmark detection methods, we can use the proposed appearance guided landmark matching module to obtain precise and stable landmark sequences, enabling the animation of any character from any driven video.

Character Animation Model. After obtaining the target landmark sequence L_{tar}, it can be applied to any image-to-video model to animate the character portrait. Specifically, L_{tar} is treated as an additional condition for the U-Net, either injected via a ControlNet-like structure (Niu et al., 2024) or incorporated directly into the latent space (Hu, 2024; Wei et al., 2024). This enables the model to precisely track the motion encoded in the landmark sequence while preserving the character's visual identity. Integrating L_{tar} at different stages of the generation process improves temporal consistency and ensures accurate expression transfer. Moreover, this flexible approach can be seamlessly extended to various architectures, enhancing scalability across diverse animation tasks.

4 EXPERIMENTS

4.1 IMPLEMENT DETAIL

As described in Section 3.2, FaceShot can be adapted to any landmark-driven animation model. In this work, we employ MOFA-Video (Niu et al., 2024), a Stable Video Diffusion (SVD)-based image animation model (Blattmann et al., 2023a), as our base model. For appearance guided landmark matching, we utilize Stable Diffusion v1.5 along with the pre-trained weights of IP-Adapter (Ye et al., 2023) to extract diffusion features from the images. Specifically, we set the time step $t = 301$, the U-Net layer $l = 6$, and the number of reference images $k = 10$.

Evaluation Metrics. Following Xie et al. (2024); Ma et al. (2024), we employ four metrics to evaluate identity similarity, high- and low-level image quality and expression accuracy. Specifically, we utilize ArcFace score (Deng et al., 2019) that calculates average cosine similarity between source and generated videos as identity similarity. We also employ Hyper-IQA (Zhang et al., 2023b) and Laion Aesthetic (Schuhmann et al., 2022) for evaluating image quality from low- and high-level, respectively. Moreover, we conduct the expression evaluation following the steps of Sun (2024): 1) Generate a local patch in the eyes & mouth areas from landmarks and create a dense grid for initial tracking points; 2) Employ CoTracker (Dinh et al., 2011) to track the generated and driving points. 3) Apply a transformation matrix from generated to driving points and normalize the trajectories based on the driving patch dimensions. 4) Compute the MSE distance for each trajectory, where lower values indicate closer alignment. We utilize the landmarks matched from our appearance guided landmark matching module in point tracking evaluation for all methods.

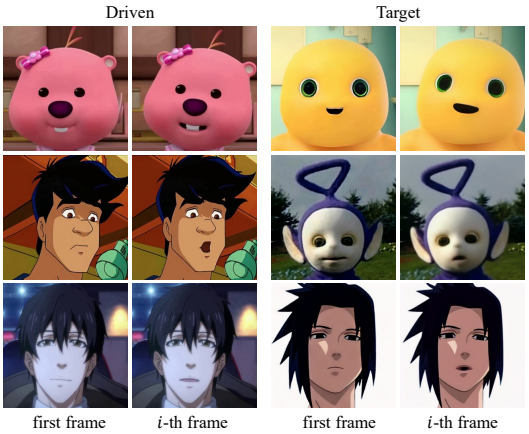

Figure 8: Visualizations of character animation from non-human driven videos.

Character Benchmark. Previous works (Ma et al., 2024; Xie et al., 2024; Yang et al., 2024) often evaluate their models on limited real-world dataset like VFHQ (Xie et al., 2022), HDTF (Zhang et al., 2021) or other personalized human portraits (Ma et al., 2024; Xie et al., 2024). To comprehensively evaluate the effectiveness and generalization ability of portrait animation methods towards characters, we build CABench that comprises 46 characters from various domains, such as animals, emojis, toys and anime characters. Characters in CABench are collected from Internet by following the guideline of ensuring that the characters do not resemble human facial features as much as possible. Moreover, CABench contains 24 driven videos of human head from RAVDESS (Livingstone & Russo, 2018), which includes 24 actors, eight emotions and two levels of emotional intensity (normal and strong). We carefully select three videos across 17 actors for each emotion. Half of the videos have normal emotional intensity, while the other half have strong emotional intensity. We conduct qualitative and quantitative comparisons on this benchmark.

4.2 COMPARISON WITH SOTA METHODS

Qualitative Results. We compare proposed FaceShot with state-of-the-art (SOTA) portrait animation methods, including MOFA-Video (Niu et al., 2024), X-Portrait (Xie et al., 2024), Face Vid2Vid (Wang et al., 2021), FADM (Zeng et al., 2023), Follow Your Emoji (Ma et al., 2024), LivePortrait (Guo et al., 2024), and MegActor (Yang et al., 2024). As AniPortrait (Wei et al., 2024) struggles with most non-human characters, we provide its quantitative results for reference only. Visual comparisons are presented in Figure 7. We observe that most methods are influenced by the prior in the driven video, resulting in human facial features appearing on character faces. In contrast, FaceShot generates vivid and stable animations for non-human characters, demonstrating the effectiveness of its robust and precise landmark generation. Furthermore, FaceShot successfully captures subtle expression movements, such as eye closure and mouth opening.

Table 1: Quantitative comparison between FaceShot and other SOTA methods on CABench. The best result is marked in **bold**, and the second-best performance is highlighted in underline. * indicates that there are some failure cases in these methods, for fair comparison, we report the values of these methods *only for reference*.

Methods	Metrics				User Preference		
	ArcFace ↑	HyperIQA ↑	Aesthetic ↑	Point-Tracking ↓	Motion ↑	Identity ↑	Overall ↑
FaceVid2Vid	0.525	33.721	4.267	**6.944**	3.58	3.83	4.52
FADM	0.633	39.402	4.522	6.993	1.93	2.04	1.96
X-Portrait	0.490	52.357	4.754	7.301	1.47	1.63	1.57
Follow Your Emoji	0.612	52.056	4.906	6.960	6.91	6.67	6.74
AniPortrait*	0.634	55.951	4.928	6.367	5.84	5.64	5.39
MegActor*	0.613	40.191	4.855	7.183	6.53	6.75	6.26
LivePortrait*	0.893	53.587	5.092	7.474	7.33	7.08	7.11
MOFA-Video	0.609	46.986	4.778	15.142	3.27	3.04	3.18
FaceShot	**0.841**	**52.979**	**5.006**	6.973	**8.14**	**8.32**	**8.27**

Also, thanks to the generalization capabilities of our appearance guided landmark matching and relative motion transfer modules, FaceShot can be employed to provide the reference landmark sequence for non-human driven videos, extending the application of portrait animation beyond human-related videos to any video. As shown in Figure 8, FaceShot can animate any character from any driven video, demonstrating its potential for open-domain portrait animations.

Quantitative Results. We conduct a quantitative comparison on the metrics mentioned in Section 4.1. Please note that some methods, such as LivePortrait, MegActor, and AniPortrait, fail to generate animations for certain characters when they are unable to detect the face. Therefore, for a fair comparison, we report the **failure rate** for these methods as follows: AniPortrait (39.13%), MegActor (36.50%), and LivePortrait (16.67%), and we calculate their metric values on successful characters only for reference purposes. Based on Table 1, FaceShot demonstrates significantly superior performance across various metrics compared to other methods on CABench. Specifically, FaceShot achieves the highest score in terms of ArcFace (0.841), demonstrating the effectiveness of the precise landmarks generated by the appearance guided landmark matching module in preserving facial identity. Additionally, FaceShot achieves superior HyperIQA (52.979) and Aesthetic (5.006) scores, indicating better image quality. Additionally, the relative landmark motion transfer module contributes to the competitive point tracking score (6.973), highlighting its ability to handle motion effectively. It is important to note that our method has achieved significant improvements across all metrics compared to the base method, MOFA-Video, further demonstrating the effectiveness of our proposed FaceShot.

User Preference. Additionally, we randomly selected 15 case examples and enlisted 20 volunteers to evaluate each method across three key dimensions: Motion, Identity, and Overall User Satisfaction. Volunteers ranked the animations based on these criteria, ensuring a fair and comprehensive comparison between the methods. As shown in Table 1, FaceShot achieves the highest scores in Motion, Identity, and Overall categories, demonstrating its robust animation capabilities across diverse characters and driven videos.

Table 2: Experiments on the number k of reference images. Best result is marked in **bold**. $\tau = 1.637$(s) represents the basic time for inference on a single image.

k	NME ↓	Time
1	12.801	τ
5	7.009	5τ
10	6.343	10τ
15	6.267	15τ
50	6.252	50τ
1000	**6.104**	1000τ

4.3 Ablation Studies

Choices of Time Step t and Layer l. To extract the diffusion feature that best fits facial instances, we conduct detailed experiments on the selection of time steps t and layer l of U-Net. Specifically, we test different combinations of t and l on 300W (Sagonas et al., 2016), a widely used facial dataset, and report Norm Mean Error (NME) as the quantitative result. As shown in Figure 9, we achieve the best NME value when $t = 301$ and $l = 6$, which are used as the basic settings of our paper.

Choice of Reference Number k. As mentioned in Section 3.2, we use the averaged diffusion feature at the i-th point of k reference images to improve matching performance. We evaluate different values of $k = 1, 5, 10, 15, 50, 1000$ on the 300W dataset and report the NME and time cost in Table 2.

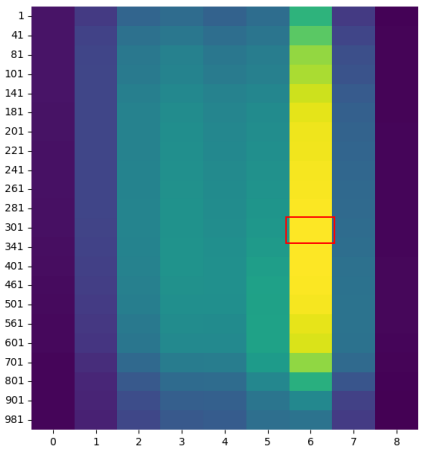

Figure 9: Heatmap of NME values for time step t and layer l of the U-Net.

Figure 10: Visual results of landmark motion transfer between 3DMM and our method.

Figure 11: Visual results of landmarks on 300W.

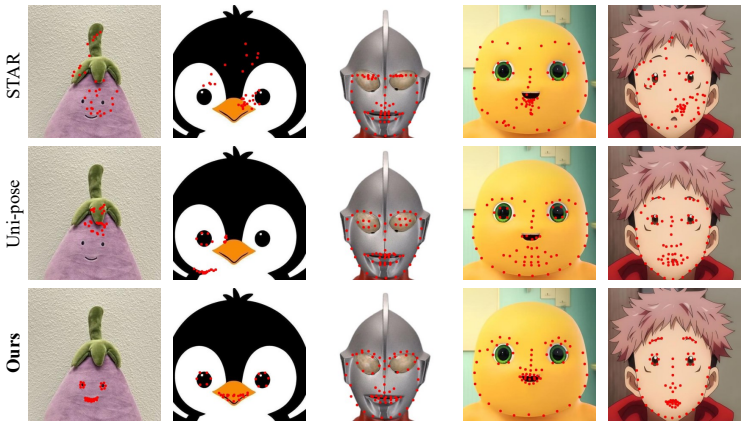

Figure 12: The visualizations of landmarks on different characters in CABench dataset using STAR, Uni-pose and our method.

Results indicate that increasing k significantly improves performance when $k \leq 10$. However, for $k > 10$, the time cost increases exponentially with diminishing performance gains. Based on this observation, we set $k = 10$ for our experiments.

Landmark Motion Transfer. To evaluate our landmark motion transfer algorithm, we provide a comparison visualization between it and 3DMM. As shown in Figure 10, our method can precisely capture the subtle motion such as mouth closure, eye closure, and even global face movement.

Appearance Guided Landmark Matching. To evaluate the effectiveness of our appearance guided landmark matching module, we first compare it with SOTA unsupervised landmark matching algorithm DIFT (Tang et al., 2023) on 300W. For a fair comparison, we also employ average feature $k = 10$ on DIFT implementation. As illustrated in Table 3, our method achieves significantly superior performance compared to DIFT. The visual landmark results on 300W are shown in Figure 11. Additionally, we provide the cosine similarity distribution during inference. As depicted in Figure 13, with prior appearance knowledge, the similarity between reference points and unrelated target points becomes smaller, reducing the probability of mismatching.

Table 3: NME values between DIFT ($k = 1, 10$) and FaceShot ($k = 10$) on 300W. Best result is marked in **bold**.

Methods	NME ↓
DIFT ($k = 1$)	18.562
DIFT ($k = 10$)	8.289
FaceShot	**6.343**

Furthermore, to verify the robust performance of our method on characters, we test it against SOTA supervised methods Uni-pose (Yang et al., 2023) and STAR (Zhou et al., 2023) on CABench. As shown in Figure 12, for human-related characters, our landmarks perfectly match the discriminative shapes of the eyes and mouth for each portrait, while other methods tend to follow human priors

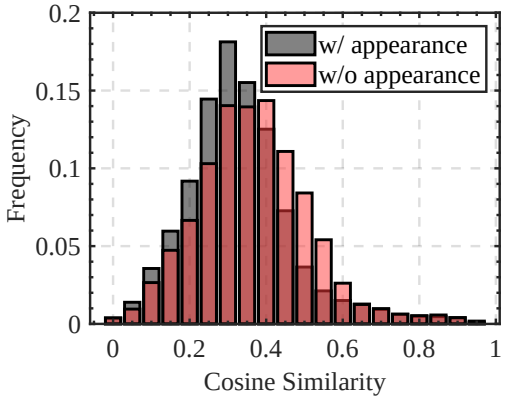

Figure 13: Cosine similarity distribution with or without appearance guidance.

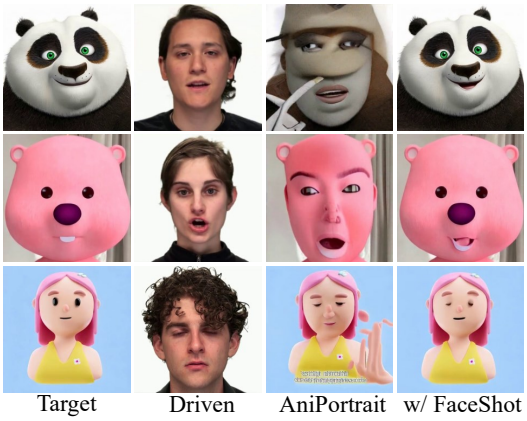

Figure 14: Visual results of AniPortrait with or without FaceShot as a plugin to generate animation results.

more closely. Moreover, our method performs well on the non-human character portraits, whereas others fail to accurately match the positions of the eyes and mouth.

As a Plugin. Experiments in Section 4.2 have demonstrated that FaceShot can improve the performance of the landmark-driven method MOFA-Video (Niu et al., 2024). To further verify the effectiveness of FaceShot as a plugin for landmark-driven animation models, we apply FaceShot to generate the target landmark results and input them into the AniPortrait (Wei et al., 2024) pipeline. As shown in Figure 14, inaccurate landmarks often result in distortions, producing generation results that resemble real humans. This occurs because landmark detection algorithms fail to generalize to non-human characters. However, FaceShot can provide precise and robust landmarks for any character, leading to harmonious and stable animation results.

5 Conclusion

In this paper, we introduced FaceShot, a training-free portrait animation framework that animates any character from any driven video. By leveraging semantic correspondence in latent diffusion model features, FaceShot addresses the limitations of existing landmark-driven methods, enabling precise landmark matching and motion transfer. This powerful capability not only extends the application of portrait animation beyond traditional boundaries but also enhances the realism and consistency of animations in landmark-driven models. FaceShot is also compatible with any landmark-driven animation model as a plugin. Additionally, we presented CABench, a standardized benchmark dataset for evaluating distribution-agnostic portrait animation methods. Experimental results demonstrate that FaceShot consistently outperforms state-of-the-art methods on CABench.

Future Work. Although FaceShot shows strong performance, future work could focus on enhancing appearance guided landmark matching by refining semantic feature extraction from latent diffusion models, particularly for complex facial geometries. Furthermore, parameterizing motion transfer could offer more precise control over facial expressions, improving the adaptability of FaceShot across diverse character types and styles.

6 Ethics Statement

In developing FaceShot, a training-free portrait animation framework that animates any characters from any driven video, we are dedicated to upholding ethical standards and promoting responsible AI use. We acknowledge potential risks, such as deepfake misuse or unauthorized media manipulation, and stress the importance of applying this technology in ways that respect privacy, consent, and individual rights. Our code will be publicly released to encourage responsible use in areas like entertainment and education, while discouraging unethical practices, including misinformation and harassment. We also advocate for continued research on safeguards and detection mechanisms to prevent misuse and ensure adherence to ethical guidelines and legal frameworks.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b.

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5543–5552, 2016.

Tim Cootes, ER Baldock, and J Graham. An introduction to active shape models. *Image processing and analysis*, 328:223–248, 2000.

Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance. *arXiv e-prints*, pp. arXiv–2311, 2023.

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*, pp. 4690–4699, 2019.

Thang Ba Dinh, Nam Vo, and Gérard Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR 2011*, pp. 1177–1184. IEEE, 2011.

Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1078–1085. IEEE, 2010.

Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2663–2671, 2022.

Litong Gong, Yiran Zhu, Weijie Li, Xiaoyang Kang, Biao Wang, Tiezheng Ge, and Bo Zheng. Atomovideo: High fidelity image-to-video generation. *arXiv preprint arXiv:2403.01800*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

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. *arXiv preprint arXiv:2307.04725*, 2023.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised keypoints from pretrained diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22820–22830, 2024.

Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3397–3406, 2022.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Xiehe Huang, Weihong Deng, Haifeng Shen, Xiubao Zhang, and Jieping Ye. Propagationnet: Propagate points to curve to learn structure information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7265–7274, 2020.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 88–97, 2017.

Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8236–8246, 2020.

Yaokun Li, Guang Tan, and Chao Gou. Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. *International Journal of Computer Vision*, 132(4): 1242–1257, 2024.

Steven R Livingstone and Frank A Russo. The ryerson 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.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.

Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024.

Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 781–790, 2018.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18444–18455, 2023.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024.

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*, pp. 8748–8763. PMLR, 2021.

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.

George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2490–2501, 2024.

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*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023.

Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47: 3–18, 2016.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*, 2023.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

Sun. Robust portrait animation. In *NA*, pp. NA, 2024.

Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3476–3483, 2013.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389, 2023.

Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024a.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10039–10049, 2021.

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. *Advances in Neural Information Processing Systems*, 36, 2024b.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2129–2138, 2018.

Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3067–3074, 2017.

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*, pp. 657–666, 2022.

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*, pp. 1–11, 2024.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *European conference on computer vision*, pp. 398–416. Springer, 2022.

Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. Unipose: Detecting any keypoints. *arXiv preprint arXiv:2310.08530*, 2023.

Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 628–637, 2023.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023a.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14071–14081, 2023b.

Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7747–7756, 2024.

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*, pp. 3661–3670, 2021.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3657–3666, 2022.

Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 386–391, 2013.

Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15475–15484, 2023.

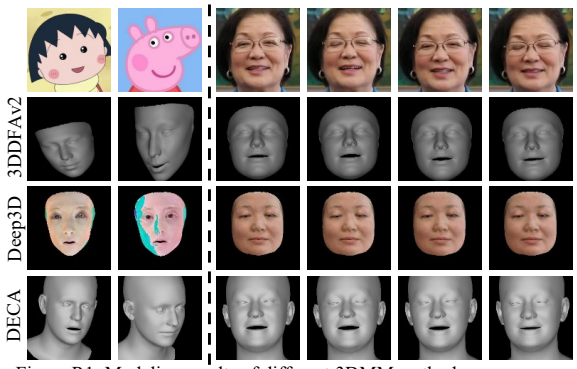

Figure R1: Modeling results of different 3DMM methods.

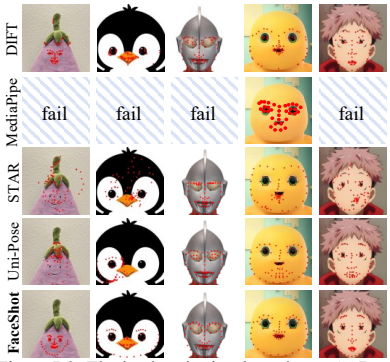

Figure R2: The landmark visual results on CABench.

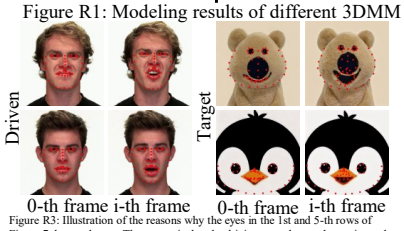

Figure R3: Illustration of the reasons why the eyes in the 1st and 5-th rows of Figure 7 do not change. The reason is that the driving eyes do not change instead of poor driving quality.

Figure R7: Texture artifacts in other diffusion-based methods.

Figure R4: Eye driving case, closed eyes, for the 1st and 5-th characters in Figure 7.

Figure R8: Illustration of why texture artifacts occur.

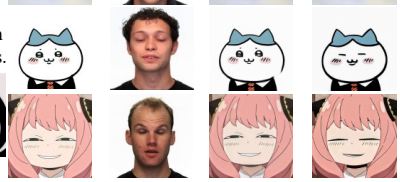

Figure R6: Comparisons with LivePortrait.

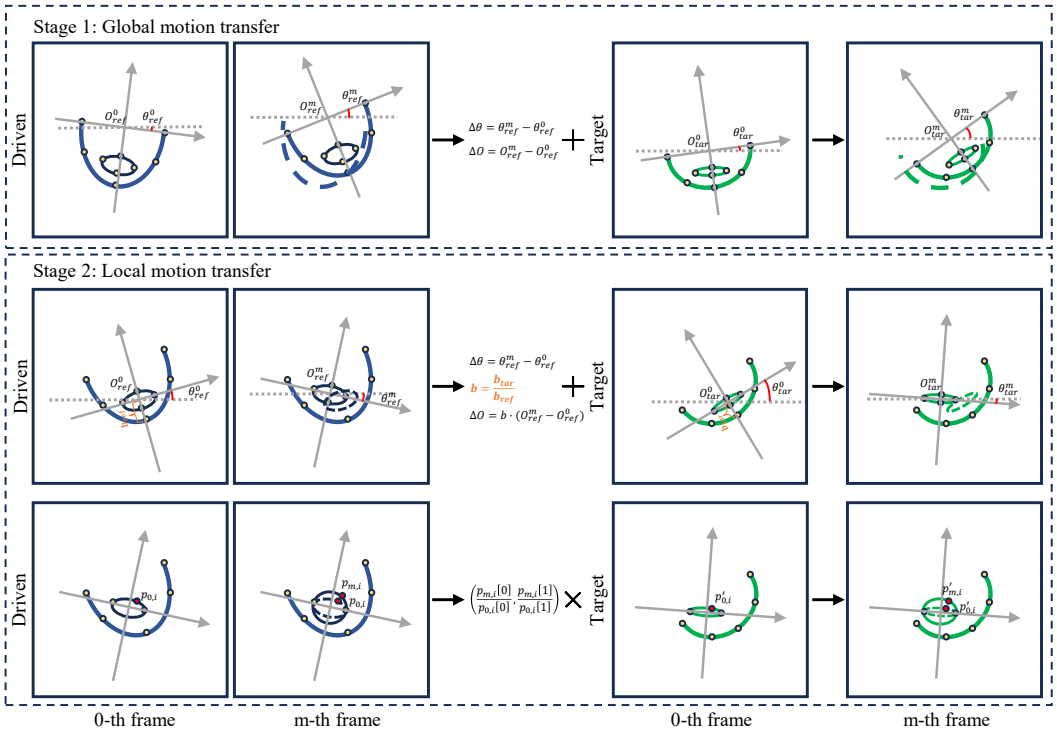

Figure R5: Illustration of our relative landmark motion transfer module.

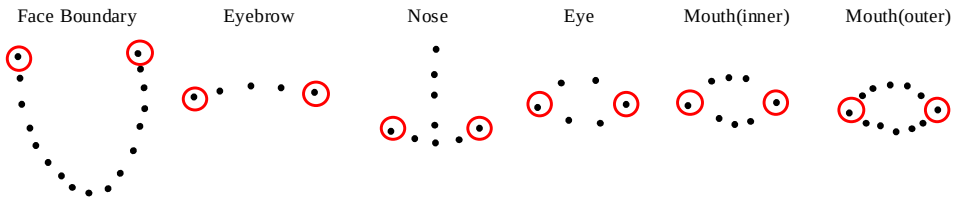

Figure R9: Illustration of the endpoints in each part, marked with a red circle.