This CVPR 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.

High-Fidelity Relightable Monocular Portrait Animation with Lighting-Controllable Video Diffusion Model

Mingtao Guo[1] Guanyu Xing[2] Yanli Liu[1,3†]
[1]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, China
[2]School of Cyber Science and Engineering, Sichuan University, Chengdu, China
[3]College of Computer Science, Sichuan University, Chengdu, China
mingtaoguo@stu.scu.edu.cn,{xingguanyu, yanliliu}@scu.edu.cn

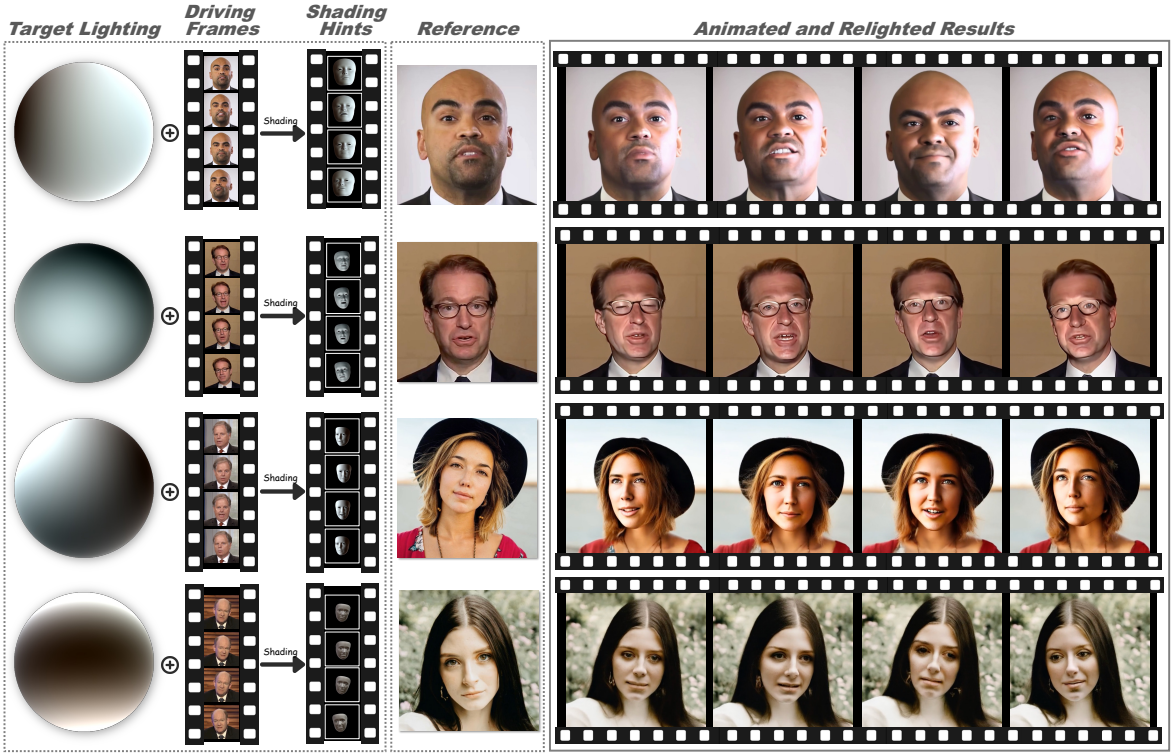


Figure 1. Qualitative results of our method. Given a reference portrait, a target lighting and a driving video, for each frame we first render a guidance image referred to as a shading hint with the 3D mesh of the reference portrait, the target lighting, and the pose of the driving frame. Then, with the shading hint, our relightable portrait animation framework relights the reference portrait during animation. The results within the solid boxes exhibit the lighting consistency with the target lighting and pose consistency with the driving frames.

Abstract

Relightable portrait animation aims to animate a static reference portrait to match the head movements of a driving video while adapting to user-specified or reference lighting conditions. Existing portrait animation methods fail to achieve relightable portraits because they do not separate and manipulate intrinsic (identity and appearance) and extrinsic (pose and lighting) features. In this paper, we present a Lighting Controllable Video Diffusion model (LCVD) for high-fidelity, relightable portrait animation. We address this limitation by distinguishing these feature types through dedicated subspaces within the feature space of a pre-trained image-to-video diffusion model. Specifically, for each frame, we use the 3D mesh of the reference portrait, the pose of the driving frame as well as the specified lighting to render an image called shading hint. While

†Corresponding author: Yanli Liu.

the reference image represents the intrinsic attributes, the shading hint encodes the extrinsic attributes. In the training phase, we employ a reference adapter to map the reference into an intrinsic feature subspace and a shading adapter to map the shading hints into an extrinsic feature subspace. By merging features from these subspaces, the model achieves nuanced control over lighting and pose in generated animations. Extensive evaluations show that LCVD outperforms state-of-the-art methods in lighting realism, image quality, and video consistency, setting a new benchmark in relightable portrait animation. Our project is available at https://github.com/MingtaoGuo/Relightable-Portrait-Animation

1. Introduction

Portrait animation has wide applications in video conferencing, virtual reality, and the film industry. With the rapid advancement of GANs [15, 24, 25] and Diffusion Models [18, 40, 46], existing portrait animation methods [11, 16, 50, 57, 59] have demonstrated remarkable capabilities in generating talking faces. For instance, the state-of-the-art LivePortrait [16] achieves real-time, high-fidelity portrait animation by designing better motion transformation and scaling up the talking head datasets. However, the ability of manipulating lighting during portrait animation remains under-explored, which is highly important for seamlessly blending the generated foreground portrait with backgrounds under varying lighting conditions.

In this paper, we focus on relightable portrait animation. We aim to animate a portrait in a still reference image, ensuring that the animated portrait's movements match those in the driving video and its lighting aligns with the user-specified conditions. From the perspective of face attributes, we reduce the task to preserving the intrinsic features (identity and appearance) of the reference portrait while effectively transferring the extrinsic ones (given pose and lighting) to the reference portrait. Obviously, the exact separation between intrinsic features and extrinsic features is crucial to reach our goal. A key reason existing portrait animation methods cannot independently manipulate lighting is their inability to decompose these two types of features.

In order to achieve high-fidelity relightable portrait animation, our key idea is to distinguish these two types of features by learning their feature subspaces, then maintain the intrinsic facial features and transfer external features. Observing that the image-to-video (I2V) diffusion model [5], trained on a large-scale dataset encompassing a variety of portraits with different lighting, poses, identities and appearances provides a foundation for learning the two feature subspaces, we design two adapters to separately model the extrinsic and intrinsic feature subspaces within the feature space of I2V. Specifically, we represent the portrait's extrinsic attributes using shading hints rendered with the reference image's 3D mesh, target lighting, and driving frame's pose, while the intrinsic attributes are represented by the reference image. During the self-supervised training process, we design a shading adapter to map the shading hints into an extrinsic feature subspace and a reference adapter to map the reference image into an intrinsic feature subspace. By merging the features from these two subspaces, the model generates portraits with specified lighting, pose, identity, and appearance.

With the I2V diffusion model, we propose a novel **L**ighting **C**ontrollable **V**ideo **D**iffusion model (**LCVD**) to achieve high-fidelity, relightable portrait animation. First, we use an off-the-shelf model [14] to extract the 3D mesh, pose, and spherical harmonics lighting coefficients for each frame of the target video. Then an image is rendered as a guidance image, referred to as a shading hint, which encodes lighting and pose information. In the training stage, to enable pose alignment and lighting control, we use a shading adapter to map the shading hints to the extrinsic feature subspace, representing external portrait attributes by establishing a mapping between the shading hints and the target portrait. For identity and appearance preservation, we use a reference adapter to map the reference image to the intrinsic feature subspace, representing internal portrait attributes by creating a mapping between an initial frame and subsequent frames. Finally, during the inference stage, we merge the features from the two subspaces and input them into the I2V diffusion model to generate portraits with specified lighting, pose, identity, and appearance. To further control the lighting magnitude, we employ multi-condition classifier-free guidance to emphasize the influence of the shading adapter and reduce the reference's impact on relighting.

Our main contributions are listed as follows:

- We introduce a Lighting Controllable Video Diffusion model, which overcomes the limitations of current portrait animation methods that fail to manipulate lighting while animating the portrait.
- We propose a shading adapter and a reference adapter to construct the feature subspaces for extrinsic and intrinsic facial features, respectively. By merging these two subspaces, the I2V model is guided to achieve relightable portrait animation.
- Extensive experiments demonstrate that LCVD surpasses state-of-the-art methods, showing significant improvements in metrics measuring lighting effects, image quality, and video consistency.

2. Related Work

2.1. Portrait Relighting

Portrait relighting adjusts the lighting of an image or video while preserving the subject's identity and appear-

ance. Previous methods [34, 47, 54] utilized One-Light-at-a-Time (OLAT) systems to capture detailed geometry and reflectance, achieving realistic relighting. However, OLAT data is expensive and difficult to acquire, limiting its practicality. To overcome this, recent approaches [20, 60] simulate multi-lighting data and train relighting networks. Despite these efforts, simulated methods still trail the realism achieved by OLAT-based techniques. Additionally, learning 3D face representations from 2D images without 3D supervision has now become feasible. Recent method [23] combines NeRF [32] with generative models [25, 40] to generate high-resolution, multi-view consistent face images.

Another simplified strategy [31, 33, 37, 39] involves capturing a selfie video or a sequence of images to obtain multi-view information. However, the rendering quality is highly dependent on the accuracy of the geometry, requiring sufficient viewpoints from the video. Additionally, these methods often need to be retrained for each new video, which makes them impractical. In contrast, our method achieves high-fidelity and temporally stable video portrait relighting, requiring only a single portrait image and a target lighting.

2.2. Diffusion-based Portrait Animation

Denoising Diffusion Models [18, 46] are based on the idea of Markov diffusion and fits the distribution of real samples by approximating a standard normal distribution. They outperform GANs [15] in sample diversity and quality, and have been successfully applied to image synthesis [3, 26, 28, 42, 55], image editing [2, 6], and video synthesis [5, 10, 51]. In portrait animation, FADM [53] refines coarsely animated portraits generated by previous methods [19, 43, 44] by combining 3DMM [4] parameters with a diffusion model for better appearance. Follow-Your-Emoji [30] instead uses expression-aware landmarks within the Animate-Anyone framework [21] to guide animation.

However, diffusion based portrait animation methods effectively transfer poses from driving images to animate the reference image but cannot simultaneously manipulate its lighting during animation.

3. Preliminaries

The Latent Diffusion Model (LDM) [40] is designed to generate high-quality, diverse images based on text prompts. It performs the denoising process within the latent space of a Variational Autoencoder (VAE) [13]. During training, the input image $\mathbf{x}_0$ is first encoded into its latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, where $\mathcal{E}(\cdot)$ represents the frozen encoder. The resulting latent code $\mathbf{z}_0$ is then perturbed as follows:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \in \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_t)$ with β_t is the noise strength at step t, and t is sampled uniformly from $\{1, \ldots, T\}$. This process is a Markov chain that adds Gaussian noise to the latent code $\mathbf{z}_0$. The denoising model ϵ_θ learns the latent space distribution by optimizing the objective function using $\mathbf{z}_t$ as input,

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}_0, \mathbf{c}, \epsilon \in \mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{2}$$

where $\mathbf{c}$ represents the condition, which is the text embedding encoded by the user-provided CLIP [38] text encoder.

4. Methodology

4.1. Overview

Our pipeline for lighting controllable portrait animation consists of two stages. First, in the training phase, we construct portrait intrinsic and extrinsic feature subspaces within a pre-trained I2V model's feature space using two adapters. Then, in the relighting and animation stage, we modify the extrinsic subspace and merge it with the intrinsic subspace to achieve relightable portrait animation, as illustrated in Fig. 2.

1. *Portrait Attributes Subspace Modeling Stage:* We employ the off-the-shelf model DECA [14] to encode each frame of the input video, extracting key parameters such as lighting, pose, and shape, which are then rendered as shading hints. After the shading hints and reference image are processed through the shading adapter and reference adapter, they are randomly selected, with each training iteration potentially including either one, both, or neither for composition (Sec. 4.2). The composed features are subsequently fed into the Stable Video Diffusion Model (SVD) [5] for self-supervised training (Sec. 4.3). The goal of this stage is to model both the extrinsic and intrinsic feature subspaces through the joint optimization of two adapters.
2. *Relighting and Animation Stage:* We render the shading hints using the pose of the portrait from the video, the shape from the reference image, and the spherical harmonics coefficients of the target lighting. Then, we combine the outputs of the shading adapter and the reference adapter to form the conditional set and employ multi-condition classifier-free guidance to adjust the magnitude of the extrinsic feature guidance direction by modifying the guidance strength, thereby generating results for lighting controllable portrait animation (Sec. 4.4).

4.2. Portrait Attributes Subspace Modeling

Current portrait animation methods can be driven by user-provided pose information. However, since portrait intrinsic and extrinsic features are entangled during the self-supervised training process, manipulating lighting requires modifying extrinsic features. This entanglement makes it difficult to adjust lighting independently during portrait animation. Therefore, separating portrait intrinsic and extrin-

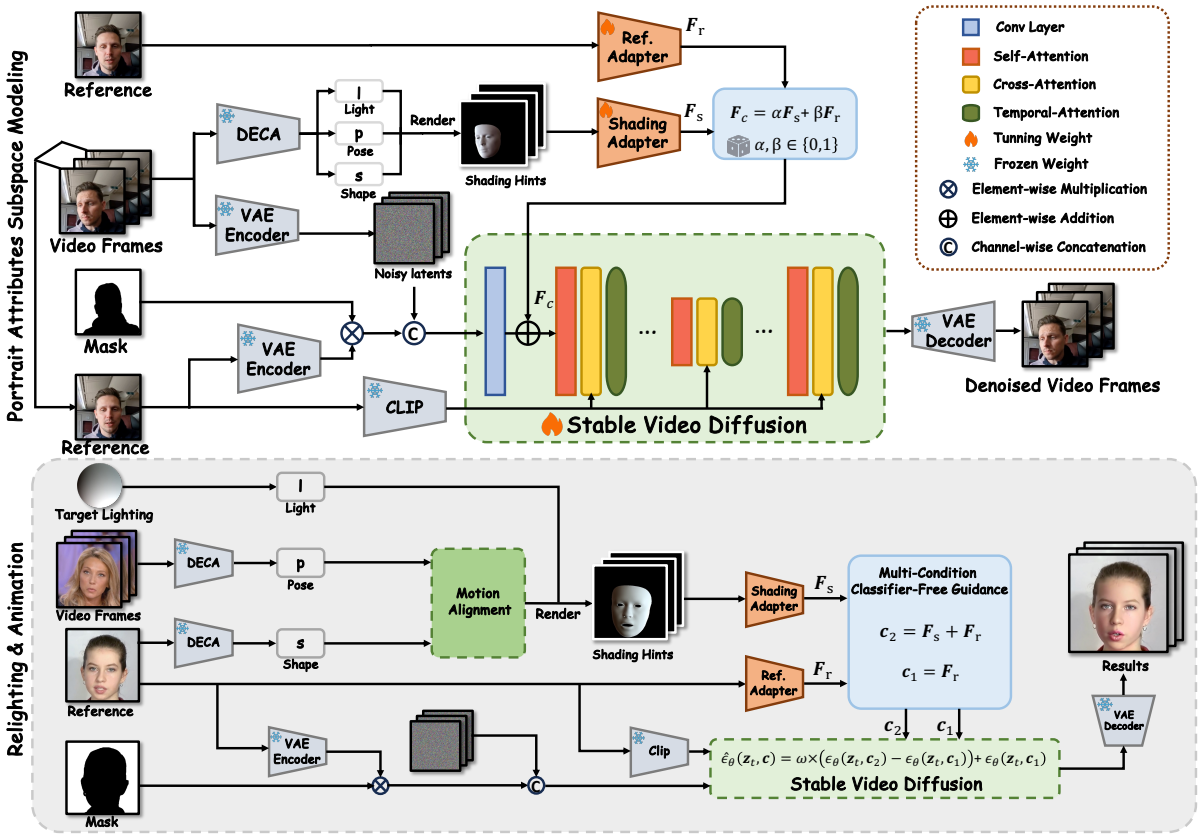


Figure 2. Overview of our pipeline for lighting controllable portrait animation. It consists of two main stages: (1) Portrait Attributes Subspace Modeling Stage: We use DECA to encode video frames and extract lighting, pose, and shape parameters, which are rendered as shading hints. After processing the shading hints and reference image through the shading adapter and reference adapter, the two features are randomly selected and fused as guidance to guide the Stable Video Diffusion Model in generating denoised video frames with consistent lighting, pose, identity, and appearance. (2) Relighting and Animation Stage: We render the shading hints using the pose of the portrait from the video, the shape from the reference image, and the spherical harmonics coefficients of the target lighting. After processing the shading hints and reference image through two adapters, we employ multi-condition classifier-free guidance to adjust the magnitude of the extrinsic feature guidance direction, enabling the generation of lighting controllable portrait animations.

sic features is a significant challenge for enabling effective lighting control in portrait animation.

To address this, we design a shading adapter and a reference adapter to construct an extrinsic feature subspace, as well as an intrinsic feature subspace within the SVD feature space during the training phase. First, we use the parametric model **FLAME** [29] as a prior to model the shape and pose attributes of human portraits.

$$\mathbf{FLAME}(s, p, e) = \mathbb{R}^{|s| \times |p| \times |e|} \rightarrow \mathbb{R}^{m \times 3}, \qquad (3)$$

which takes shape coefficients $s \in \mathbb{R}^{|s|}$, pose $p \in \mathbb{R}^{|p|}$, and expression $e \in \mathbb{R}^{|e|}$ as inputs to generate the corresponding 3D face mesh. We use DECA to estimate these parameters, with the added benefit of DECA's ability to predict second-order spherical harmonics lighting coefficients $l \in \mathbb{R}^{|l|}$. We then render the 3D face mesh with spherical harmonics to produce a lighting-shaded face, called shading hints.

We process the input video fragment into a sequence of shading hints, which, together with the reference image, are independently transformed by the shading and reference adapters into features $\mathbf{F}_s$ and $\mathbf{F}_r$. To establish these two feature subspaces, $\mathbf{F}_s$ and $\mathbf{F}_r$ are recombined with random coefficients $\{\alpha, \beta | \alpha, \beta \in \{0, 1\}\}$ and input into the SVD. This method enables the SVD to explore both extrinsic and intrinsic feature subspaces in its feature space.

4.3. Lighting-Guided Video Diffusion Model

We choose the SVD [5] as the prior model for our LCVD method. However, SVD is an image-guided video generation model that takes an input image $I \in \mathbb{R}^{H \times W \times 3}$. This image is first encoded by CLIP's vision encoder [38] and passed into SVD's cross-attention module. At the same time, the image is encoded by a Variational Autoencoder (VAE) [13] into a latent representation $\mathbf{z}_0 = \mathcal{E}(I) \in \mathbb{R}^{h \times w \times c}$. This latent $\mathbf{z}_0$ is then replicated T times and concatenated along the channel dimension with noise $\hat{\mathbf{z}} \in \mathbb{R}^{T \times h \times w \times c}$, resulting in $\mathbf{z}_t \in \mathbb{R}^{T \times h \times w \times 2c}$. The resulting $\mathbf{z}_t$ is then input into a 3D UNet [41], which progressively

denoises the input to generate a video of T frames. Here, we set $h = H/8$, $w = W/8$, and $c = 4$.

For instance, when an image of a dog walking on the street is input into the SVD, the model predicts the next T frames of the dog based on the original image. As a result, the subsequent frames generated by SVD inherit the objects and lighting conditions from the original image. To eliminate the influence of the original image's lighting on the relighting results, as shown in Fig. 2, we use a mask to remove the portrait from the reference image's latent space.

During the training phase, we use a mask $\mathcal{M}$ to remove the portrait, compensating for the loss of identity and appearance information using the reference adapter. Additionally, we incorporate each frame's portrait mask into the loss function, encouraging the model to focus more on the portrait region. The loss function is defined as follows:

$$\mathcal{L}_p = \mathbb{E}\left[\| \left(1 - \mathcal{M}\right)\left(\epsilon - \epsilon_\theta\left(\mathbf{z}_t, t, \mathbf{c}\right)\right) \|\right], \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, I) \in \mathbb{R}^{T\times h\times w\times c}$, and the portrait mask $\mathcal{M} \in \{0,1\}^{T\times h\times w\times c}$. Finally, the total loss is:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_{LDM}. \tag{5}$$

4.4. Lighting Controllable Portrait Animation

In the relighting and animation stage, we incorporate the reference image, video fragment, and target lighting. When the portrait in the reference image corresponds to the same individual as that in the video fragment, we utilize DECA to extract the pose information from the video portrait and the shape information from the reference image, subsequently rendering a sequence of shading hints based on the lighting coefficients derived from the target lighting. However, in cases where the portraits in the video and reference image do not represent the same individual, we introduce a motion alignment module to mitigate the risk of identity leakage from the video portrait, which could compromise the quality of the generated output (for further details of motion alignment, please refer to the supplementary materials).

Following this, we input the shading hints and the reference image into the shading adapter and reference adapter. Given that the portrait in the reference image inherently contains its own lighting information, directly combining features may result in the original lighting dominating the shading hints, leading to ineffective relighting. To solve this, we adopt the concept of Composer [22]. This allows us to achieve portrait lighting manipulation by adjusting the direction of the lighting guidance within the set of conditions. The formula is as follows:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) = \omega\left(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_2) - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_1)\right) + \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_1), \tag{6}$$

here, $\mathbf{c}_1$ and $\mathbf{c}_2$ are two sets of conditions. If a condition exists in $\mathbf{c}_2$ but not in $\mathbf{c}_1$, its strength is enhanced by a weight ω. The larger ω, the stronger the condition. If a condition exists in both $\mathbf{c}_1$ and $\mathbf{c}_2$, ω has no effect, and the condition strength defaults to 1.0. In this way, we can set $\mathbf{c}_2 = \mathbf{F}_s + \mathbf{F}_r$ and $\mathbf{c}_1 = \mathbf{F}_r$, where $\mathbf{F}_s$ and $\mathbf{F}_r$ are the features from the shading adapter and reference adapter. Since $\mathbf{F}_s$ is present in $\mathbf{c}_2$ but not in $\mathbf{c}_1$, we can enhance the strength of the extrinsic feature by adjusting ω. At the same time, because both $\mathbf{c}_1$ and $\mathbf{c}_2$ contain $\mathbf{F}_r$, the reference image's portrait features remain intact. This allows us to achieve relightable portrait animation by utilizing classifier-free guidance for condition combination.

5. Experiments

5.1. Implementation Details

Datasets. We train our model on the CelebV-HQ [61] and VFHQ [49] datasets. Since the backbone of SVD [5] is sensitive to video quality, we first evaluate each video in two datasets with the video quality assessment method Faster-VQA [48], and remove videos with scores lower than 0.6. In the end, 37,644 videos remain for training. To ensure a fair comparison in experiments, we evaluate our method on the portrait video dataset HDTF [58] and FFHQ [24].

Training Details. During the training phase, for the temporal attention layers of the SVD, we sample 16-frame video sequences to establish temporal consistency, with each frame at a resolution of 512×512. Unlike methods such as [21, 30], which require two separate training stages, we update all the weights of both the SVD and two adapters simultaneously. The model is trained for 30,000 steps with a batch size of 8 using gradient accumulation, optimized by 8bit-Adam [27] with a learning rate of 1×10^{-5}.

Inference Details. During the inference stage, we employ DDIM [46] with 25 sampling steps and set the default guidance weight ω to 4.5. For a 100-frame video, this approach requires about 2 minutes and 10 GB of VRAM for inference on an NVIDIA 4090 GPU.

5.2. Metrics and Comparisons

Evaluation Metrics. To evaluate the performance of our method, following [8], we relight the first 100 frames of each video in the HDTF dataset. Each video is rendered with four distinct lighting conditions derived from four different lighting-effect reference faces, resulting in a total of 44,000 frames for a comprehensive comparison. Following [23], we use an off-the-shelf estimator [14] to calculate the Lighting Error (LE). Arcface [12] is used to measure Identity Preservation (ID) between the relit results and the original images. To assess temporal consistency, we compute LPIPS [56] between adjacent frames. We further employ an image quality assessment model [9] and a video quality assessment model [48] to evaluate Image Quality (IQ) and Video Quality (VQ), respectively. Additionally, Fréchet Inception Distance (FID) [17] and Fréchet Video Distance

(FVD) [45] are used to measure video fidelity. Besides objective evaluation, we conduct a user study where 17 participants rate the videos based on three criteria: Lighting Accuracy (LA-User), Identity Similarity (ID-User), and Video Quality (VQ-User). Each criterion is rated on a 1 to 5 scale: poor, fair, average, good, and excellent. Finally, we calculate the average score for each criterion across participants.

Comparative Methods. For the portrait relighting task, we conduct a comparative analysis between LCVD and five state-of-the-art portrait relighting methods: DPR [60], SMFR [20], NFL [23], StyleFlow [1], and DiFaReli [35], evaluating performance on both the HDTF and FFHQ datasets. For the portrait animation task, we compare LCVD with three state-of-the-art portrait animation methods: DaGAN [19], StyleHEAT [52], and AnimateAnyone [21], using the HDTF dataset for evaluation.

5.3. Quantitative Evaluation

In portrait video relighting, Table 1 shows that our method outperforms other state-of-the-art methods in all metrics except for ID. Specifically, it improves video fidelity (FVD) by 32%, image fidelity (FID) by 16%, and image quality (IQ) by 14.6% compared to the second-best method, demonstrating excellent video quality. While our method does not achieve the highest ID performance, this is because relighting in our method is applied during portrait animation, where ID information is derived only from the reference, unlike other methods that relight each frame individually. However, our method achieves the best ID performance in the user study, likely due to its higher-quality, more stable video synthesis, which visually aligns with better ID preservation. This also demonstrates the ID loss in our method is within an acceptable range for perception.

Since NFL [23], StyleFlow [1], and DiFaReli [35] are trained on the aligned FFHQ facial dataset, we compare our method on 500 FFHQ images for a fair evaluation. As shown in Table 2, our method outperforms the second-best method in identity preservation (ID) by 11.4% and image quality (IQ) by 16.1%. However, it does not achieve the best performance in lighting error (LE) and image fidelity (FID) because these methods are trained on FFHQ, while our model is trained on different video datasets, resulting in slightly lower lighting and fidelity performance. Notably, as our method is for video sequences and FFHQ is an image dataset, we replicate each image 16 times to form a video sequence for image testing adaptation.

In addition to portrait relighting, we use the lighting and shape from the reference image and the pose from the driving image to render shading hints, guiding our model to achieve cross-identity portrait animation, which we then evaluate. Beyond the previously mentioned metrics, we incorporate a POSE metric to assess the pose accuracy of the animated portraits, ensuring alignment with the poses in the driving video. The POSE evaluation method follows that of [44], using a facial landmark detection model [7] to measure the pose error between the animated portraits and the driving portraits based on facial keypoints. As shown in Table 3, our method outperforms the other methods in all metrics, particularly achieving a 29.7% improvement in image fidelity (FID), a 10.1% improvement in image quality (IQ), and an 8.7% improvement in identity preservation (ID) compared to the second-best method.

5.4. Qualitative Evaluation

We compare our approach with previous portrait relighting methods on the HDTF dataset, including state-of-the-art face alignment-based approaches such as StyleFlow [1], NFL [23], and DiFaReli [35]. Additionally, we compare our method with face alignment-free methods like DPR [60] and SMFR [20]. The results are shown in Fig. 3. We find that face alignment-based methods easily suffer from background detail loss and identity degradation, especially in pre-trained StyleGAN-based [25] methods like StyleFlow and NFL (e.g., see the results in the fourth and fifth columns, where the background details are completely lost, and the facial identity is inconsistent with the input). On the other hand, DiFaReli, based on a pre-trained diffusion model [36], benefits from the DDIM inverse [46] method. It successfully reconstructs background details and preserves identity, but introduces noticeable face artifacts.

Although face alignment-free methods like DPR and SMFR achieve relighting without losing background and facial identity, the trade-off is a significant reduction in image quality, with the lighting appearing unnatural, as if a shadow has been cast over the image (e.g., in the first and second rows of the third column for SMFR). In contrast, our method in the final column greatly outperforms others in both image quality and the realism of the lighting effects. Notably, our approach accurately renders specular reflections on the face and eyes, as well as realistic shadows cast by facial muscles, while keeping identity loss within acceptable limits. The background details are also preserved. Overall, our approach demonstrates superior capability.

Since NFL, StyleFlow, and DiFaReli are trained on the aligned FFHQ dataset, we visualize the relighting results on FFHQ for a fair comparison. As shown in Fig. 4, NFL and StyleFlow lose background details and alter the portrait identity. DiFaReli preserves background details but introduces facial artifacts, lowering image quality. In contrast, our method maintains background details and identity consistency, achieving optimal image quality.

Additionally, we compare our method with DaGAN, StyleHEAT, and AnimateAnyone for portrait animation. As shown in Fig. 5, while DaGAN preserves the pose from the driving frame, the portrait identity differs significantly from the reference, and the image quality is low. StyleHEAT in-

Table 1. Quantitative comparison of portrait relighting with DPR, SMFR, NFL, StyleFlow, and DiFaReli based on objective evaluation and user study on the HDTF video dataset. The best scores are highlighted in bold, and the second-best are underlined.

	Objective Evaluation							User Study		
Methods	LE↓	ID↑	LPIPS↓	IQ↑	VQ↑	FID↓	FVD↓	LA-User↑	ID-User↑	VQ-User↑
DPR [60]	0.768	**0.730**	0.0295	2.646	0.734	44.57	403.0	3.423	3.462	3.125
SMFR [20]	0.747	0.601	0.0333	1.057	0.588	60.50	551.6	3.047	2.877	2.604
NFL [23]	0.784	0.199	0.0823	2.586	0.766	96.17	819.3	2.894	2.553	2.398
StyleFlow [1]	0.932	0.474	0.1088	2.614	0.746	161.3	900.6	2.103	1.929	1.563
DiFaReli [35]	0.783	0.531	0.1152	1.103	0.458	57.49	743.2	3.141	2.592	2.284
Ours	**0.738**	0.585	**0.0282**	**3.034**	**0.775**	**37.46**	**273.3**	**3.534**	**4.000**	**3.398**

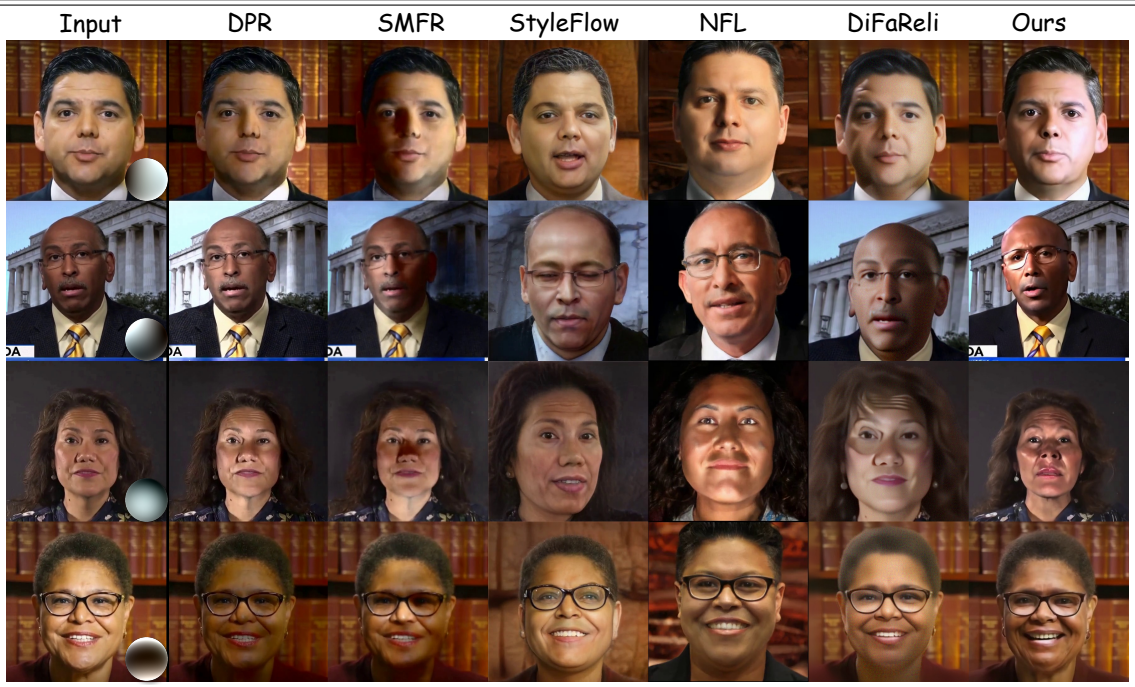


Figure 3. Qualitative comparisons with DPR [60], SMFR [20], StyleFlow [1], NFL [23], and DiFaReli [35]. The first column shows the input video frames, and the remaining columns present relighted results under various lighting conditions. Our method demonstrates more realistic performance, particularly in challenging cases such as side lighting.

Table 2. Quantitative comparison of portrait relighting with NFL, StyleFlow and DiFaReli on the FFHQ dataset. The best scores are highlighted in bold, and the second-best are underlined.

Methods	LE↓	ID↑	IQ↑	FID↓
NFL[23]	0.892	0.253	3.020	118.9
StyleFlow[1]	1.042	0.485	3.846	102.7
DiFaReli[35]	**0.749**	0.687	1.591	**25.98**
Ours	0.938	**0.765**	**4.465**	26.71

troduces distortions in cross-identity portrait animation, and although AnimateAnyone, a diffusion model guided by a reference-net, generates higher image quality, it still suffers from identity loss and occasional facial artifacts.

5.5. Ablation Study

Effectiveness of Adapters. Our method constructs intrinsic and extrinsic feature subspaces using the reference and

Table 3. Quantitative comparison of cross-identity portrait animation with DaGAN, StyleHEAT, and AnimateAnyone on the HDTF dataset. The best scores are highlighted in bold, and the second-best scores are underlined.

Methods	ID↑	POSE↓	IQ↑	VQ↑	FID↓
DaGAN[19]	0.645	3.935	1.005	0.528	107.4
StyHE.[52]	0.201	34.58	1.554	0.612	149.9
AniAny.[21]	0.806	5.086	2.744	0.706	69.85
Ours	**0.876**	**3.805**	**3.021**	**0.717**	**49.11**

shading adapters, respectively, enabling relightable portrait animation by merging these subspaces. We conduct an ablation study with different adapter combinations. First, when retaining only the shading adapter as shown in Fig. 6, the column labeled F_s illustrates that the generated portrait's pose and lighting align with the shading hints, indicating that only the extrinsic features are transferred. When only the reference adapter is used, the column labeled F_r shows



Figure 4. Qualitative comparison of portrait relighting with NFL [23], StyleFlow [1], and DiFaReli [35] on the FFHQ dataset [24]. The first column shows the input FFHQ portrait images, and the remaining column display the relighted results under various lighting conditions. Our method demonstrates more realistic results.



Figure 5. Qualitative comparison of cross-identity portrait animation with DaGAN [19], StyleHEAT [52], and AnimateAnyone [21] on the HDTF dataset. Our method yields more lifelike results.

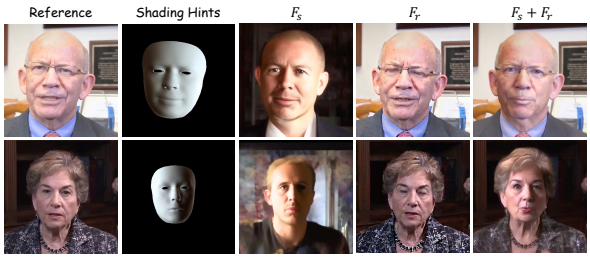


Figure 6. Ablation study comparing the performance of our model in portrait generation under different adapter combinations. F_s uses only the shading adapter, F_r uses only the reference adapter, and $F_s + F_r$ uses both.

that the generated portrait closely resembles the reference with only minor variations, such as blinking, indicating intrinsic feature preservation. When both adapters are used, the column labeled $F_s + F_r$ demonstrates that the generated portrait not only matches the pose and lighting of the shading hints but also maintains the identity and appearance of the reference.



Figure 7. Ablation study comparing our model with varying strengths of multi-condition classifier-free guidance ω. As ω increases, the relighting effect increasingly aligns with the target lighting; however, this comes at the cost of some loss of identity information and a decrease in image quality.

Table 4. Quantitative comparison of ablation study with portrait mask in the training phase. "w/" indicates that the mask is used, and "w/o" indicates that the mask is not used. The best scores are highlighted in bold, and the second-best are underlined.

Metrics	LE↓	IQ↑	Metrics	LE↓	IQ↑
w/o mask	1.090	1.781	w/ mask	**0.738**	**3.034**

Effectiveness of Guidance Strength. In Fig. 7, we visualize the relighting results for different ω values. When $\omega = 2$, the lighting effect is minimal, with only small differences from the input image, resulting in good identity retention. In contrast, when $\omega = 8$, the lighting effect closely aligns with the target lighting, but this also leads to reduced image quality and some loss of identity retention. The primary reason for this phenomenon is that as ω increases, the proportion of extrinsic features grows, while the proportion of intrinsic features diminishes, resulting in a degradation of identity information from the reference image. Consequently, higher values of ω enhance lighting effects but lead to greater identity loss.

Effectiveness of Portrait Mask. During the training phase, we remove the portrait mask input (w/o mask) and test the performance on the HDTF dataset. As Table 4 shows, w/ mask has greater advantages than w/o mask in LE and IQ. LE drops by 32.3% and IQ rises by 70.4%, indicating the portrait mask is effective in lighting and image quality.

6. Conclusion

In this paper, we introduce a Lighting Controllable Video Diffusion model (LCVD) for high-fidelity, relightable portrait animation. By distinguishing between intrinsic and extrinsic facial features, the model effectively preserves identity and appearance while enabling precise control over lighting and pose. The model also incorporates multi-condition classifier-free guidance to fine-tune the lighting effects. Our extensive experimental results demonstrate that LCVD significantly outperforms existing methods in lighting realism, image quality, and video consistency.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Project (No.62472297, No.62172290), Sichuan Science and Technology Program (No.2024ZHCG0007). We would like to thank the anonymous reviewers for their helpful comments.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3), 2021. 6, 7, 8

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 3

[3] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 3

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 3

[5] 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*, 2023. 2, 3, 4, 5

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[7] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 6

[8] Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6221–6231, 2024. 5

[9] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 5

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 3

[11] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 2

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3, 4

[14] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 2, 3, 5

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 3

[16] 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. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[19] 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*, pages 3397–3406, 2022. 3, 6, 7, 8

[20] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14719–14728, 2021. 3, 6, 7

[21] 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*, pages 8153–8163, 2024. 3, 5, 6, 7, 8

[22] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *International Conference on Machine Learning*, pages 13753–13773. PMLR, 2023. 5

[23] Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. Nerffacelighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM Transactions on Graphics (TOG)*, 2023. 3, 5, 6, 7, 8

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 5, 8

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3, 6

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[27] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 4

[30] 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. 3, 5

[31] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4263–4273, 2024. 3

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[33] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 3

[34] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4):43–1, 2021. 3

[35] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22646–22657, 2023. 6, 7, 8

[36] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[37] Haonan Qiu, Zhaoxi Chen, Yuming Jiang, Hang Zhou, Xiangyu Fan, Lei Yang, Wayne Wu, and Ziwei Liu. Relitalk: Relightable talking portrait generation from a single video. *International Journal of Computer Vision*, pages 1–16, 2024. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4

[39] Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun BR, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, et al. Lite2relight: 3d-aware single image portrait relighting. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[41] 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*, pages 234–241. Springer, 2015. 4

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[43] 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. 3

[44] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 3, 6

[45] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022. 6

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 5, 6

[47] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[48] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5

[49] 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*, pages 657–666, 2022. 5

[50] 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*, pages 1–11, 2024. 2

[51] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[52] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 6, 7, 8

[53] 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*, pages 628–637, 2023. 3

[54] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 802–812, 2021. 3

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[57] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2

[58] 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*, pages 3661–3670, 2021. 5

[59] 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*, pages 3657–3666, 2022. 2

[60] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 3, 6, 7

[61] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 5