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

Supplementary Material

7. Ablation Study

Effectiveness of Adapters. The shading adapter maps shading hints to the extrinsic feature subspace, while the reference adapter maps the reference to the intrinsic feature subspace. The combination of features from these different subspaces enables various effects, such as controlling lighting magnitude, maintaining identity, and enhancing image generation quality. To investigate the effectiveness of these adapters, we conducted an ablation study with different adapter combinations.

First, we retain only the reference adapter, as shown in Table 5 under the row F_r. In this case, the lighting error is significant (LE is large), while identity preservation is excellent (ID is high). This indicates that the model preserves intrinsic features well but fails to capture extrinsic features. Conversely, when we retain only the shading adapter, as shown in the row F_s, the lighting error is minimal (LE is small), but identity preservation is almost nonexistent (ID approaches 0). This suggests that the model transfers extrinsic features effectively while neglecting intrinsic features.

When both adapters are retained, we observe significant improvements in intrinsic feature preservation compared to using only the shading adapter and significant improvements in extrinsic feature transfer compared to using only the reference adapter. Moreover, the image quality also achieves its optimal level under this configuration.

Effectiveness of Guidance Strength. This method utilizes a multi-condition classifier-free guidance approach to control the lighting magnitude through the classifier-free guidance mechanism [18]. The strength of the guidance, represented by ω, directly affects the lighting intensity.

To evaluate the impact of ω, we conduct an ablation study with varying values, as shown in Table 6. As ω increases, the lighting effect improves (LE decreases), but identity preservation deteriorates (ID decreases). Notably, image quality reaches its peak at $\omega = 4$. However, setting ω too high can lead to a decline in image quality. Therefore, lighting effects, identity preservation, and image quality can be balanced by appropriately adjusting the value of ω.

Table 5. Quantitative comparison of ablation study with different adapter combinations on the HDTF dataset. F_r denotes using only the reference adapter, F_s denotes using only the shading adapter, and $F_s + F_r$ represents using both adapters. The best scores are highlighted in bold, and the second-best are underlined.

Methods	LE↓	ID↑	IQ↑	FID↓
F_r	1.071	**0.802**	<u>1.662</u>	**35.61**
F_s	**0.582**	0.028	1.248	56.63
$F_s + F_r$	<u>0.738</u>	<u>0.585</u>	**3.034**	<u>37.46</u>

Table 6. Quantitative comparison of the ablation study on the impact of different guidance strengths ω on lighting (LE), identity (ID), and image quality (IQ) on the HDTF dataset. From left to right, each metric is shown as it changes with increasing ω. The best scores are highlighted in bold, and the second-best are underlined.

Methods	$\omega = 2$	$\omega = 4$	$\omega = 6$	$\omega = 8$
LE↓	1.079	0.809	<u>0.744</u>	**0.681**
ID↑	**0.728**	<u>0.603</u>	0.563	0.503
IQ↑	2.611	**2.988**	<u>2.954</u>	2.856

8. Motion Alignment

As shown in Fig. 2, during the relighting and animation stages, we use a video to animate the reference image, ensuring that the lighting effect of the relit portrait is consistent with that of the target lighting. In the inference stage, since the portrait in the video and the reference image come from different identities, directly using the shading hints of the portrait from the video to animate the reference image would cause the generated portrait to resemble the one from the driving video. This leads to identity leakage during animation, degrading the animation quality. We propose two motion alignment methods: (1) a relative displacement-based motion alignment method and (2) a portrait scale consistency-based motion alignment method.

Relative Displacement-based Motion Alignment. This motion alignment method is designed to use the reference image as the first frame, with subsequent motions based on this initial frame. The motion guidance for the reference frame is achieved by leveraging the relative displacement between consecutive frames in the driving video. First, we use DECA to extract the pose sequence $\mathbf{P} = \{p_1^v, p_2^v, \ldots, p_n^v\}$ and the expression sequence $\mathbf{E} = \{e_1^v, e_2^v, \ldots, e_n^v\}$ from each frame of the driving video, along with the pose p^R and shape s^R from the reference image. Next, we calculate the relative pose offsets $\Delta\mathbf{P} = \{0, p_2^v - p_1^v, \ldots, p_n^v - p_1^v\}$ for each frame with respect to the first frame. Using the reference image's pose p^R as the base pose, we then apply these relative offsets to obtain an aligned pose sequence $\mathbf{P}^{align} = \{p^R, p^R + (p_2^v - p_1^v), \ldots, p^R + (p_n^v - p_1^v)\}$. Finally, we combine the expression sequence $\mathbf{E}$ with the reference image's shape s^R and the aligned pose sequence $\mathbf{P}^{align}$. These parameters are

then input into Eq. 3 to obtain $\mathbf{FLAME}(s^R, \mathbf{P}^{align}, \mathbf{E})$, which, along with the spherical harmonic lighting coefficients l from the target lighting, is used to render the shading hints for each frame.

Portrait Scale Consistency-based Motion Alignment. The relative displacement-based alignment method relies on using the reference image as the base frame for relative motion. However, this approach does not ensure perfect spatial alignment between the pose of the generated portrait and the driving video. To address this, we propose an alternative motion alignment method aimed at achieving perfect alignment between the generated portrait's pose and that of the driving video. Specifically, we first use DECA to extract the pose sequence $\mathbf{P} = \{p_1^v, p_2^v, \ldots, p_n^v\}$ and the expression sequence $\mathbf{E} = \{e_1^v, e_2^v, \ldots, e_n^v\}$ from each frame of the driving video, along with the shape s^R from the reference image. These parameters are then input into Eq. 3 to compute $\mathbf{FLAME}(s^R, \mathbf{P}, \mathbf{E})$. Combined with the spherical harmonic lighting coefficients l from the target lighting, this process renders the shading hints for each frame.

9. Shading and Reference Adapter Network Architecture

As shown in Fig. 8, the network architecture of the shading adapter and reference adapter is illustrated. These two networks map shading hints and the reference image into the *extrinsic feature subspace* and *intrinsic feature subspace* of SVD's feature space, respectively. As depicted in Fig. 2, the two features are fused with the features from the first convolutional layer of SVD. Therefore, the shading hints and reference image must match the spatial dimensions and channel count of the output from SVD's first convolutional layer. To achieve this, we designed the network structure shown in Fig. 8.

Moreover, since SVD is designed for video sequence generation, the output dimensions of its first layer include an additional temporal dimension F, resulting in an output shape of $B \times F \times C \times H \times W$. Accordingly, the input to the shading adapter is a sequence of shading hints with dimensions $B \times F \times C \times H \times W$. For the reference image, which consists of a single frame with dimensions $B \times 1 \times C \times H \times W$, we duplicate the reference F times to obtain dimensions $B \times F \times C \times H \times W$ before feeding it into the reference adapter.

10. Long Video Sequence Generation

Since our model is based on the SVD backbone, which is limited to generating video sequences of 16 frames at a time, we tackle the challenge of animating portrait videos of arbitrary length by utilizing the diffusion model sampling method proposed in [59]. To ensure smooth transitions between consecutive video segments, we implement

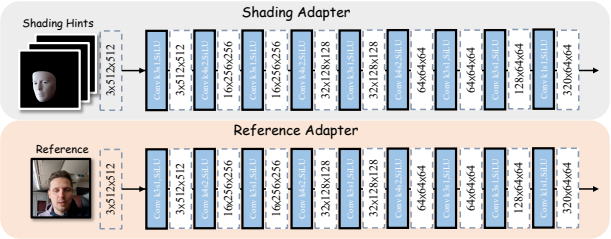

Figure 8. Network architecture of the shading adapter and reference adapter, where k denotes the kernel size and s denotes the stride. These two networks have the same structure but do not share weights and are updated alongside SVD during the training phase.

a 6-frame overlap strategy. In our experiments, we employ DDIM with 25 sampling steps and set the default guidance weight ω to 4.5. For a 100-frame video, this method takes approximately two minutes and 10 GB of VRAM to perform inference on an NVIDIA 4090 GPU.