Diffusion Video Autoencoders: Toward Temporally Consistent Face Video Editing via Disentangled Video Encoding

Gyeongman Kim[1] Hajin Shim[1] Hyunsu Kim[2] Yunjey Choi[2] Junho Kim[2] Eunho Yang[1,3]
[1]Korea Advanced Institute of Science and Technology (KAIST), South Korea
[2]NAVER AI Lab [3]AITRICS, South Korea
{gmkim, shimazing, eunhoy}@kaist.ac.kr {hyunsu1125.kim, yunjey.choi, jhkim.ai}@navercorp.com

Figure 1. **Face video editing**. Our editing method shows improvement compared to the baseline [35] in terms of temporal consistency (left, "eyeglasses") and robustness to the unusual case such as the hand-occluded face (right, "beard").

Abstract

*Inspired by the impressive performance of recent face image editing methods, several studies have been naturally proposed to extend these methods to the face video editing task. One of the main challenges here is temporal consistency among edited frames, which is still unresolved. To this end, we propose a novel face video editing framework based on diffusion autoencoders that can successfully extract the decomposed features - for the first time as a face video editing model - of identity and motion from a given video. This modeling allows us to edit the video by simply manipulating the temporally invariant feature to the desired direction for the consistency. Another unique strength of our model is that, since our model is based on diffusion models, it can satisfy both reconstruction and edit capabilities at the same time, and is robust to corner cases in wild face videos (*e.g. *occluded faces) unlike the existing GAN-based methods.*[1]

1. Introduction

As one of the standard tasks in computer vision to change various face attributes such as hair color, gender, or glasses of a given face image, face editing has been continuously gaining attention due to its various applications and entertainment. In particular, with the improvement of analysis and manipulation techniques for recent Generative Adversarial Network (GAN) models [8, 11, 22, 29, 30], we simply can do this task by manipulating a given image's latent feature. In addition, very recently, many methods for face image editing also have been proposed based on Diffusion Probabilistic Model (DPM)-based methods that show high-quality and flexible manipulation performance [3, 12, 17, 19, 23, 25].

Naturally, further studies [2, 35, 41] have been proposed to extend image editing methods to incorporate the temporal axis for videos. Now, given real videos with a human face, these studies try to manipulate some target facial attributes with the other remaining features and motion intact. They all basically edit each frame of a video independently via off-the-shelf StyleGAN-based image editing techniques [22, 29, 41].

Despite the advantages of StyleGAN in this task such as high-resolution image generation capability and highly disentangled semantic representation space, one harmful drawback of GAN-based editing methods is that the encoded real

[1]Project page: https://diff-video-ae.github.io

images cannot perfectly be recovered by the pretrained generator [1, 26, 34]. Especially, if a face in a given image is unusually decorated or occluded by some objects, the fixed generator cannot synthesize it. For perfect reconstruction, several methods [6, 27] are suggested to further tune the generator for GAN-inversion [26, 34] on one or a few target images, which is computationally expensive. Moreover, after the fine-tuning, the original editability of GANs cannot be guaranteed. This risk could be worse in video domain since we have to finetune the model on the multiple frames.

Aside from the reconstruction issue of existing GAN-based methods, it is critical in video editing tasks to consider the temporal consistency among edited frames to produce realistic results. To address this, some prior works rely on the smoothness of the latent trajectory of the original frames [35] or smoothen the latent features directly [2] by simply taking the same editing step for all frames. However, smoothness does not ensure temporal consistency. Rather, the same editing step can make different results for different frames because it can be unintentionally entangled with irrelevant motion features. For example, in the middle row of Fig. 1, eyeglasses vary across time and sometimes diminish when the man closes his eyes.

In this paper, we propose a novel video editing framework for human face video, termed *Diffusion Video Autoencoder*, that resolves the limitations of the prior works. First, instead of GAN-based editing methods suffering from imperfect reconstruction quality, we newly introduce diffusion based model for face video editing tasks. As the recently proposed *diffusion autoencoder* (DiffAE) [23] does, thanks to the expressive latent spaces of the same size as input space, our model learns a semantically meaningful latent space that can perfectly recover the original image back and are directly editable. Not only that, for the first time as a video editing model, encode the decomposed features of the video: 1) identity feature shared by all frames, 2) feature of motion or facial expression in each frame, and 3) background feature that could not have high-level representation due to large variances. Then, for consistent editing, we simply manipulate a single invariant feature for the desired attribute (single editing operation per video), which is also computationally beneficial compared to the prior works that require editing the latent features of all frames.

We experimentally demonstrate that our model appropriately decomposes videos into time-invariant and per-frame variant features and can provide temporally consistent manipulation. Specifically, we explore two ways of manipulation. The first one is to edit features in the predefined set of attributes by moving semantic features to the target direction found by learning linear classifier in semantic representation space on annotated CelebA-HQ dataset [10]. Additionally, we explore the text-based editing method that optimizes a time-invariant latent feature with CLIP loss [7]. It is worth noting that since we cannot fully generate edited images for CLIP loss due to the computational cost, we propose the novel strategy that rather uses latent state of intermediate time step for the efficiency.

To summarize, our contribution is four-fold:

- We devise diffusion video autoencoders based on diffusion autoencoders [23] that decompose the video into a single time-invariant and per-frame time-variant features for temporally consistent editing.
- Based on the decomposed representation of diffusion video autoencoder, face video editing can be conducted by editing only the single time-invariant identity feature and decoding it together with the remaining original features.
- Owing to the nearly-perfect reconstruction ability of diffusion models, our framework can be utilized to edit exceptional cases such that a face is partially occluded by some objects as well as usual cases.
- In addition to the existing predefined attributes editing method, we propose a text-based identity editing method based on the local directional CLIP loss [7, 24] for the intermediately generated product of diffusion video autoencoders.

2. Related Work

Video editing When editing a given real video, it is essential to preserve temporal consistency. First, Lai *et al*. [14] consider editing global features of videos such as artistic style transfer, colorization, image enhancement, etc. To encourage temporal consistency, they train sequential image translator with temporal loss defined as the warping error between the output frames.

Different from this work, our target task is to edit facial attributes of human face video. In other words, we aim to change face-related attributes such as eyeglasses, beards, etc. For this task, Yao *et al*. [41] are the first that propose the editing pipeline: 1) align and crop the target face area, 2) encode these frames to latent features, manipulate and decode them, and 3) unalign and paste the edited frames to the original video. However, every step of this pipeline is conducted independently for each frame. For the manipulation step, they try to find disentangled editing directions for a desired attribute and take the same step for every frame, while expecting the disentanglement brings consistency automatically. However, the results show inconsistency and we conjecture that both latent space and learned direction are still entangled with many other attributes. Next, Tzaban *et al*. [35] additionally fine-tune the pretrained StyleGAN [27] to enhance reconstructability. Meanwhile, they assume that temporal consistency would be preserved when applying the same editing step to the latent features of the original frames. However, their assumption is not always true for all attributes, especially for beards and eyeglasses as shown in Fig. 1. Alauf *et al*. [2] also try to avoid temporal inconsistency by smoothing latent features.

While the methods in the previous paragraph handle inconsistency implicitly, several works [39, 40] try to solve it more directly. After per-frame editing, Xu *et al*. [40] further optimize latent codes and the pretrained StyleGAN to enhance consistency between the frame pairs defined by involving bi-directional optical flow. On the other hand, Xia *et al*. [39] propose to learn the dynamics of inverted GAN latent codes of video frames. The learned dynamics are used to generate the subsequent latent features after editing only the first frame.

Additionally, although Skorokhodov *et al*. [31] mainly target to generate video, they also demonstrate manipulation results of the generated video. They model the video with a time-invariant content code and per-frame motion codes. Based on this modeling, the generated video can be manipulated with a target text by optimizing its content vector with CLIP loss. However, they cannot edit the real videos due to the absence of an encoding method. In contrast, our proposed method is capable of encoding and manipulating realistic face videos.

Diffusion models Denoising diffusion probabilistic models (DDPMs) [9] associate image generation with the sequential denoising process of isotropic Gaussian noise. The model is trained to predict the noise from the input image. Unlike other generative models such as GANs and most traditional-style VAEs that encode input data in a low-dimensional space, diffusion models have a latent space that is the same size as the input. Although DDPMs require a lot of feed-forward steps to generate samples, their image fidelity and diversity are superior to other types of generative models. Compared to DDPMs that assume a Markovian noise-injecting forward diffusion process, Denoising diffusion implicit models (DDIMs) [33] assume a non-Markovian forward process that has the same marginal distribution as DDPMs, and use its corresponding reverse denoising process for sampling, which enables acceleration of the rather onerous sampling process of DDPMs. DDIMs also utilize a deterministic forward-backward process and therefore show nearly-perfect reconstruction ability, which is not the case for DDPMs. In this paper, we adopt conditional DDIMs to encode, manipulate and decode real videos.

Image editing with diffusion models There are various attempts to manipulate images with diffusion models [3, 12, 17, 19, 23, 25] such as text-guided inpainting [3, 19, 25], stroke-based editing [17], and style transfer [12]. Among those, we are especially interested in facial attribute editing tasks. Kim *et al*. [12] propose text-based image manipulation by optimizing an unconditional diffusion model to minimize local directional CLIP loss [7]. This method is superior to GAN-based methods as it can successfully edit even occluded or overly decorated faces by the excellent reconstruction ability of diffusion models. However, due to the absence of semantically meaningful latent space, there exist some semantic features that are unable to be changed. Another parallel work proposes diffusion autoencoder (DiffAE) [23], which takes a learnable encoder to obtain semantic representations of which the underlying diffusion model is conditioned. The model can perform face attribute manipulation by moving the semantic vector to a target direction. Based on DiffAEs, we design our diffusion video autoencoders to perform temporally-consistent video editing.

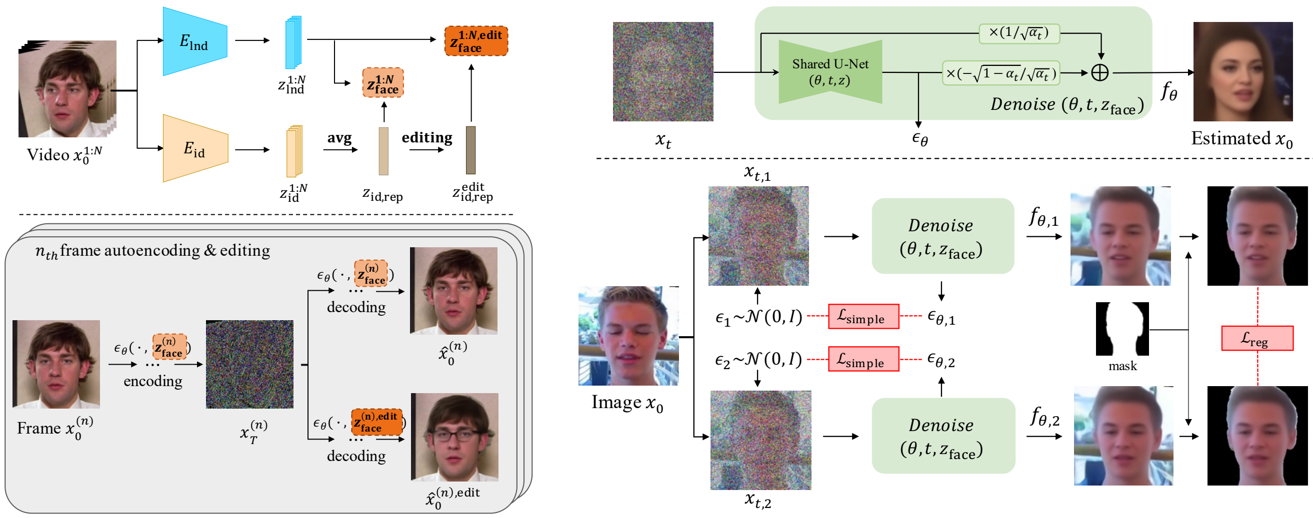
3. Preliminaries

Diffusion probabilistic models (DPMs) DPMs [9, 32] are generative models that attempt to approximate the data distribution $q(x_0)$ via $p_\theta(x_0)$ from the reverse prediction of Markovian diffusion process $q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$. Here, the forward process of DPM is a Gaussian noise perturbation $q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ where β_t is fixed or learned variance schedule with increasing $\beta_1 < \beta_2 < \cdots < \beta_T$, adding gradually increasing noise to data x_0, and its reverse process is denoising of x_t step by step in reverse order. From this definition, a noisy image x_t at step t can be expressed as $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon,\ \epsilon \sim \mathcal{N}(0, I)$ where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$. While the true reverse process $q(x_{t-1}|x_t)$ is intractable, diffusion models approximate $q(x_{t-1}|x_t, x_0)$ with $p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \sigma_t)$ by minimizing the variational lower bound of negative log-likelihood. Finally, by reparameterizing μ_θ, the objective is given as $\mathbb{E}_{x_0\sim q(x_0), \epsilon_t\sim\mathcal{N}(0,I), t}\left\|\epsilon_\theta(x_t, t) - \epsilon_t\right\|_2^2$ where the parameterized model $\epsilon_\theta(x_t, t)$ estimates the true Gaussian noise term $\frac{1}{\sqrt{1-\alpha_t}}(x_t - \sqrt{\alpha_t}x_0)$ in the forward process.

Speeding up a time-consuming sampling process of diffusion models is one of the core research topics related to diffusion models [20, 33]. Among others, Song *et al*. [33] propose DDIM, which assumes a non-Markovian forward process $q(x_t|x_{t-1}, x_0)$ that has the same marginal distribution $q(x_t|x_0)$ with DDPMs. The corresponding generative process of DDIM is $x_{t-1} = \sqrt{\alpha_{t-1}}f_\theta(x_t, t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t^2 z$ where $z \sim \mathcal{N}(0, I)$, σ is sampling stochasticity and $f_\theta(x_t, t)$, the estimated value of x_0 at time step t, is as follow:

$$f_\theta(x_t, t) = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}.$$

When σ is set to 0, this process becomes deterministic. With deterministic forward and reverse processes, we can obtain



(a) Video autoencoding and editing pipeline.

(b) Training objective.

Figure 2. Overview of our Diffusion Video Autoencoder.

the latent code of each original sample, which can reconstruct the original sample through the reverse process.

Diffusion autoencoders Due to the determinacy of its sampling process, DDIM can reconstruct the original image x_0 from x_T obtained through T forward process steps. While x_T can be considered as a latent state with the same size as x_0, it does not contain high-level semantic information [23]. To supplement this, Prechakul *et al*. propose diffusion autoencoder (DiffAE) [23] based on DDIM. DiffAE utilizes two forms of latent variables: $z_{\texttt{sem}}$ for the useful high-level semantic representation and x_T for the remaining low-level stochasticity information. DiffAE introduces a semantic encoder $\texttt{Enc}(x_0)$ which extracts $z_{\texttt{sem}}$ from an image, and makes a noise estimator $\epsilon_\theta(x, t, z_{\texttt{sem}})$ conditioned on $z_{\texttt{sem}}$. Stochastic latent x_T is calculated through the deterministic DDIM forward process [33] that involves the noise estimator ϵ_θ given $z_{\texttt{sem}}$. Then, $(z_{\texttt{sem}}, x_T)$ is decoded to reconstruct the corresponding original image x_0 with $p_\theta(x_{0:T}|z_{\texttt{sem}}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, z_{\texttt{sem}})$. DiffAE is trained with simple DDPM loss, as done with DDIM. The differences between DiffAE and DDIM are that an additional variable $z_{\texttt{sem}}$ is involved as a conditioning variable during the reverse process (and thus the training process) and that the noise estimator ϵ_θ and semantic encoder $\texttt{Enc}$ are jointly trained. As a result, through the training phase, $z_{\texttt{sem}}$ is learned to capture the high-level semantic information of the image, and the low-level stochastic variations remain in x_T.

4. Diffusion Video Autoencoders

In this section, we introduce a video autoencoder, called *Diffusion Video Autoencoder*, specially designed for face video editing to have 1) superb reconstruction performance and 2) an editable representation space for the identity feature of the video disentangled with the per-frame features changing over time. The details of the model components and the training procedure are explained in Sec. 4.1 and then two different editing methods based on our model are presented in Sec. 4.2. We provide the overview in Fig. 2.

4.1. Disentangled Video Encoding

To encode the video with N frames $\{x_0^{(n)}\}_{n=1}^{N}$, we consider the time-invariant feature of human face videos as human identity information, and the time-dependent feature for each frame as motion and background information. Among these three, identity or motion information relevant to a face is appropriate to be projected to a low-dimensional space to extract high-level representation. Comparatively, the background shows high variance with arbitrary details and changes more with the head movement by cropping and aligning the face region. Therefore, it could be very difficult to encode the background information into a high-level semantic space. Thus, identity and motion features are encoded in high-level semantic space $z_{\texttt{face}}^{(n)}$, combining identity feature $z_{\texttt{id}}$ of the video and motion feature $z_{\texttt{lnd}}^{(n)}$ of each frame, and the background feature is encoded in noise map $x_T^{(n)}$ (see Fig. 2a). We denote $z_{\texttt{id}}$ without superscript (n) for frame index since it is the time-invariant and shared across all frames of the video.

To achieve this decomposition, our model consists of two separated semantic encoders - an identity encoder $E_{\texttt{id}}$ and a landmark encoder $E_{\texttt{lnd}}$ - and a conditional noise estimator ϵ_θ for diffusion modeling. The encoded features $(z_{\texttt{id}}, z_{\texttt{lnd}}^{(n)})$ by the two encoders are concatenated and passed through an MLP to finally get a face-related feature $z_{\texttt{face}}^{(n)}$. Next, to encode noise map $x_T^{(n)}$, we run the deterministic forward process of DDIM using noise estimator ϵ_θ with conditioned on $z^{(n)}_{\texttt{face}}$. Since noise map $x^{(n)}_T$ is a spatial variable with the same size as the image, it is expected that information in the background can be encoded more easily without loss of spatial information. Then, the encoded features $(x^{(n)}_T, z^{(n)}_{\texttt{face}})$ can be reconstructed to the original frame by running generative reverse process of conditional DDIM in a deterministic manner:

$$p_\theta(x_{0:T}|z_{\texttt{face}}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, z_{\texttt{face}}).$$

To obtain the identity feature disentangled with the motion, we choose to leverage a pretrained model for identity detection, named ArcFace [5]. ArcFace is trained to classify human identity in face images regardless of poses or expressions, so we expect it to provide the disentangled property we need. Nevertheless, when an identity feature is extracted for each frame through the pretrained identity encoder, the feature may be slightly different for each frame because some frames may have partial identity features for some reasons (*e.g.* excessive side view poses). To alleviate this issue, we average the identity features $z^{(n)}_{\texttt{id}} = E_{\texttt{id}}(x^{(n)}_0)$ of all frames in the inference phase. Similarly, we obtain per-frame motion information via a pretrained landmark detection model [4] which outputs the position of face landmarks. Several studies [21,36] have shown that it is possible to have a sufficiently disentangled nature by extracting features through a pretrained encoder without learning. Therefore, diffusion video autoencoders extract an identity and also a landmark feature of the image through the pretrained encoders and map them together to the high-level semantic space for face features through an additional learnable MLP.

Next, we explain how the learnable part of our model is trained. Fig. 2b summarizes our training process. For simplicity, from now on, we drop the superscript of frame index. Our objective consists of two parts. The first one is the simple version of DDPM loss [9] as

$$\mathcal{L}_{\texttt{simple}} = \mathbb{E}_{x_0\sim q(x_0), \epsilon_t\sim\mathcal{N}(0,I), t}\left\|\epsilon_\theta(x_t, t, z_{\texttt{face}}) - \epsilon_t\right\|_1$$

where $z_{\texttt{face}}$ is an encoded high-level feature of input image x_0. It encourages the useful information of the image to be well contained in the semantic latent $z_{\texttt{face}}$ and exploited by ϵ_θ for denoising. Secondly, we devise a regularization loss to hinder face information (identity and motion) from leaking to x_T but contained in $z_{\texttt{face}}$ as much as possible for clear decomposition between background and face information. If some face information is lost in $z_{\texttt{face}}$, the lost information would remain in the noise latent x_T unintentionally. To avoid it, we sample two different Gaussian noises ϵ_1 and ϵ_2 to obtain different noisy samples $x_{t,1}$ and $x_{t,2}$, respectively. Then, we minimize the difference between the estimated original images $f_{\theta,1}$ and $f_{\theta,2}$ except for the background part as:

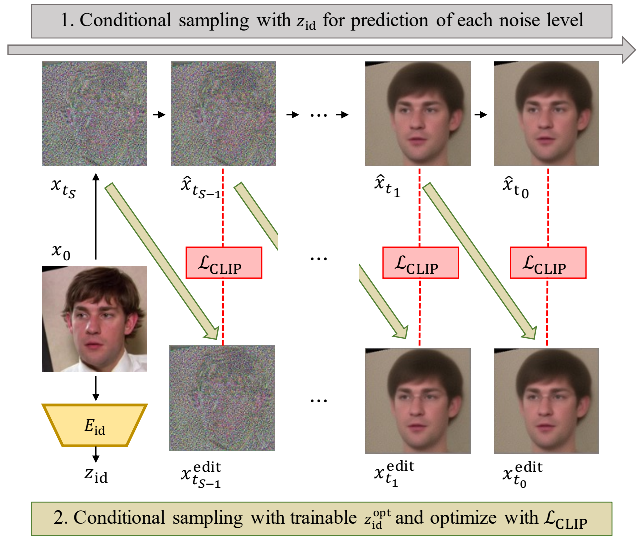


Figure 3. Process of computing noisy CLIP loss.

$$\mathcal{L}_{\texttt{reg}} = \mathbb{E}_{x_0\sim q(x_0), \epsilon_1,\epsilon_2\sim\mathcal{N}(0,I), t}\left\|f_{\theta,1}\odot m - f_{\theta,2}\odot m\right\|_1$$

where m is a segmentation mask of a face region in the original image x_0 and $f_{\theta,i} = f_\theta(x_{t,i}, t, z_{\texttt{face}})$. As a result, the effect of noise in x_t on the face region will be reduced and $z_{\texttt{face}}$ will be responsible for face features for the generation (See Fig. 7). In Sec. 5.5, we demonstrate the desired effect of $\mathcal{L}_{\texttt{reg}}$. The final loss of diffusion video autoencoders is as follows: $\mathcal{L}_{\texttt{dva}} = \mathcal{L}_{\texttt{simple}} + \mathcal{L}_{\texttt{reg}}$.

4.2. Video Editing Framework

We now describe our video editing framework with our diffusion video autoencoders. First, all video frames $\{I_n\}_{n=1}^N$ are aligned and cropped for interesting face regions as in [35]. The cropped frames $\{x^{(n)}_0\}_{n=1}^N$ are then encoded to latent features using our diffusion video autoencoder. To extract the representative identity feature of the video, we average identity features of each frame as

$$z_{\texttt{id,rep}} = \frac{1}{N}\sum_{n=1}^{N}(z^{(n)}_{\texttt{id}}),$$

where $z^{(n)}_{\texttt{id}} = E_{\texttt{id}}(x^{(n)}_0)$. Similarly, the per-frame landmark feature is computed as $z^{(n)}_{\texttt{lnd}} = E_{\texttt{lnd}}(x^{(n)}_0)$ to finally obtain per-frame face features $z^{(n)}_{\texttt{face}} = MLP(z_{\texttt{id,rep}}, z^{(n)}_{\texttt{lnd}})$. After that, we compute $x^{(n)}_T$ with DDIM forward process conditioning $z^{(n)}_{\texttt{face}}$. Thereafter, manipulation is conducted by editing $z_{\texttt{id,rep}}$ to $z^{\texttt{edit}}_{\texttt{id, rep}}$ using a pretrained linear attribute-classifier of the identity space or text-based optimization which we will discuss in more detail later. After modifying the representative identity feature $z^{\texttt{edit}}_{\texttt{id,rep}}$, the edited frame $\hat{x}^{(n),\texttt{edit}}_0$ is generated by the conditional

Figure 4. Comparison of temporal consistency to the previous video editing methods for 'beard'.

reverse process with $(z^{(n),\texttt{edit}}_{\texttt{face}}, x^{(n)}_T)$ where $z^{(n),\texttt{edit}}_{\texttt{face}} = MLP(z^{\texttt{edit}}_{\texttt{id,rep}}, z^{(n)}_{\texttt{lnd}})$. Afterward, as with all previous work, the face part of the edited frame is pasted to the corresponding area of the original frame to create a clean final result. For this process, we segment face regions using a pretrained segmentation network [42]. Below, we explore two editing methods for our video editing framework.

Classifier-based editing First, as in DiffAE [23], we train a linear classifier $C_{\texttt{attr}}(z_{\texttt{id}}) = \texttt{sigmoid}(w^{\top}_{\texttt{attr}} z_{\texttt{id}})$ for each attribute $\texttt{attr}$ on CelebA-HQ's [10] with its attribute annotation in the identity feature space. To change $\texttt{attr}$, we move the identity feature $z_{\texttt{id}}$ to $\ell_2\texttt{Norm}(z_{\texttt{id}} + s w_{\texttt{attr}})$ with a scale hyperparameter s.

CLIP-based editing Since the pretrained classifier allows editing only for several predefined attributes, we additionally devise the CLIP-guidance identity feature optimization method. Directional CLIP loss [7] requires two images corresponding to one for neutral text and one for target text, respectively. It implies that we need the synthesized images with our diffusion model, which is costly with full generative process. Therefore, to reduce the computational cost, we use the drastically reduced number of steps S $(\ll T)$ for image sampling. In other words, we consider the time steps $t_1, t_2, \ldots, t_S$ where $0 = t_0 < t_1 < \ldots < t_S \leq T$ and evenly split T. Thereafter, we compute x_{t_S} from the given image x_0 that we want to edit (normally chosen as the first frame $x^{(1)}_0$ of the video) through S-step of forward process with $z_{\texttt{id}} = E_{\texttt{id}}(x_0)$. Through the sequential reverse steps from x_{t_S}, we recover $\hat{x}_{t_s}$ for each time t_s where $s = (S-1), ..., 0$ with the original $z_{\texttt{id}}$. Meanwhile, $x^{\texttt{edit}}_{t_s}$ are obtained by the single reverse step from $\hat{x}_{t_{s+1}}$ but with variable $z^{\texttt{opt}}_{\texttt{id}}$ being optimized initialized to $z_{\texttt{id}}$ (See Fig. 3). Finally, we minimize the directional CLIP loss [7] between intermediate images $\hat{x}_{t_s}$ and $x^{\texttt{edit}}_{t_s}$, which are still noisy, for all s with a neutral (*e.g.* "face") and a target text (*e.g.* "face with eyeglasses"). We choose intermediate images $x^{\texttt{edit}}_{t_s}$, $\hat{x}_{t_s}$ from $\hat{x}_{t_{s+1}}$ instead of estimated x_0 to compute CLIP loss because to estimate x_0 directly from x_{t_s} is incomplete and erroneous for large t_s. Therefore, we expect that conservatively choosing intermediate images will help to find the edit direction more reliably. We refer the reader to Appendix B for the ablation study of our noisy CLIP loss. In addition to the CLIP loss, to preserve the remaining attributes, we also use ID loss (cosine distance between $z_{\texttt{id}}$ and $z^{\texttt{opt}}_{\texttt{id}}$) and ℓ_1 loss between face parts of $\hat{x}_{t_s}$ and $x^{\texttt{edit}}_{t_s}$ for all s. After all, the learned editing step $\Delta z_{\texttt{id}} = z^{\texttt{opt}}_{\texttt{id}} - z_{\texttt{id}}$ is applied to $z_{\texttt{id,rep}}$.

Table 1. **Quantitative reconstruction results** on the randomly chosen 20 videos in VoxCeleb1 test set. The reported values are the mean of the averaged per-frame measurements for each video.

Method	SSIM ↑	MS-SSIM ↑	LPIPS ↓	MSE ↓
e4e [34]	0.509	0.761	0.157	0.037
PTI [27]	0.765	0.939	0.063	0.007
Ours ($T = 20$)	0.540	0.905	0.228	0.016
Ours ($T = 100$)	**0.922**	**0.989**	**0.045**	**0.002**

Table 2. **Quantitative results** to evaluate temporal consistency. Ours show the best global coherency and comparable local consistency to the baselines.

Method	TL-ID	TG-ID
Yao *et al*. [41]	0.989	0.920
Tzaban *et al*. [35]	0.997	0.961
Xu *et al*. [40]	1.002	0.983
Ours	0.995	0.996

5. Experiments

In this section, we present the experimental results to confirm reconstruction performance (Sec. 5.1), temporally consistent editing ability (Sec. 5.2), robustness to the unusual samples (Sec. 5.3), and disentanglement of the encoded features (Sec. 5.4) with further ablation study (Sec. 5.5). For this purpose, we train our diffusion video autoencoder on 77294 videos of the VoxCeleb1 dataset [18]. As preprocessing, frames of the video are aligned and cropped as in [35] like the FFHQ dataset. Next, they are resized to the size of 256^2. For the noise estimator ϵ_θ, the UNet improved by [20] is used, and we train diffusion video autoencoder for 1 million steps with a batch size of 16. Please see more details in Appendix A.

5.1. Reconstruction

For video editing, the ability to reconstruct the original video from the encoded one must be preceded. Otherwise, we will lose the original one before we even edit the video. To compare this ability with baselines quantitatively, we measure frequently used metrics for reconstruction - SSIM [37], Multi-scaled (MS) SSIM [38], LPIPS [43], and MSE - on randomly selected 20 videos in the test set of VoxCeleb1. As baselines, we compared our model with the GAN-based inversion method e4e [34] and PTI [27] used by Yao *et al*. [41] and Tzaban *et al*. [35], respectively. Since StyleGAN-based methods handle the higher resolution images with the size of 1024^2, we resize the reconstructed results to 256^2 for comparison. For our method, we vary the number of diffusion steps T to observe computational cost and image quality trade-offs. In Tab. 1, our diffusion video autoencoder with $T = 100$ shows the best reconstruction ability and still outperforms e4e with only $T = 20$.

Input Yao *et al*. [41] Tzaban *et al*. [35] Ours

Figure 5. Editing wild face videos that GAN-based prior works struggled with. Classifier-based editing is used to make the person "young" (up) or change a "gender" (below).

5.2. Temporal Consistency

In Fig. 4, a visual comparison is conducted to evaluate the video editing performance of our diffusion video autoencoder qualitatively. We edit the given video to generate a beard through text-based guidance except for Yao *et al*. [41], which only allows to edit predefined attributes, by moving to the opposite direction of "no_beard". As a result, we demonstrate that only our diffusion video autoencoder successfully produces the consistent result. Specifically, Yao *et al*. [41] fail to preserve the original identity due to the limitations of GAN inversion. In the result of Tzaban *et al*. [35], the shape and the amount of beard constantly changes according to the lip motion. Although Xu *et al*. [40] show better but not perfect consistency, the motions unintentionally change as a side effect. The inconsistency can be observed more clearly between the second and the fifth column except for ours showing a reliable result.

Furthermore, we quantitatively evaluate the temporal consistency of Fig. 4 in Tab. 2. Although there is no perfect metric for temporal consistency of videos, TL-ID and TG-ID [35] imply the local and global consistency of identity between adjacent frames and across all frames, respectively, compared to the original. These metrics can be interpreted as being consistent as the original is when their values are close to 1. We emphasize that we greatly improve global consistency. TL-ID of Xu *et al*. [40] is larger than 1 because the motion of the editing results shrinks so that the adjacent frames become closer to each other than the original. Additional quantitative comparisons of temporal consistency, editing quality, and time cost are summarized in Appendix C.

5.3. Editing Wild Face Videos

Owing to the reconstructability of diffusion models, editing wild videos that are difficult to inversion by GAN-based methods becomes possible. As shown in Fig. 5, unlike others, our method robustly reconstructs and edits the given images effectively.

Figure 6. Demonstration of the disentangled video encoding.

Figure 7. Ablation of using regularization loss for training. Regularization loss helps the model to contain identity and motion information in $z_{\texttt{face}}$ as much as possible.

5.4. Decomposed Features Analysis

To demonstrate whether the diffusion video autoencoder decomposes the features adequately, we examine the synthesized images by changing each element of the decomposed features. To this end, we encode the frames of two different videos and then generate samples with a random noise or exchange the respective elements with each other in Fig. 6. When we decode the semantic latent $z_{\texttt{face}}$ with a Gaussian noise instead of the original noise latent x_T, it has a blurry background that is different from the original one, while identity and head pose are preserved considerably. This result implies that x_T contains only the background information, as we intended. Moreover, the generated images with switched identity, motion, and background feature confirm that the features are properly decomposed and the diffusion video autoencoder can produce a realistic image even with the new combination of the features. However, with extreme changes in the head position, the background occluded by the face is not properly generated because the taken x_T has no information about background in that area (See the last column of the lower example in Fig. 6).

5.5. Ablation Study

We further conduct an ablation study on our proposed regularization loss $\mathcal{L}_{\texttt{reg}}$. First, we train the model with the same setting but ablating $\mathcal{L}_{\texttt{reg}}$. As shown in Fig. 7, neither model has a problem with reconstruction. However, without the regularization loss, the identity changes significantly according to the random noise. Thus we can conclude that the regularization loss helps the model to decompose features effectively.

6. Conclusions

To tackle the temporal consistency problem in editing human face video, we have proposed a novel framework with the newly designed video diffusion autoencoder which encodes the identity, motion, and background information in a disentangled manner and decodes after editing a single identity feature. Through the disentanglement, the most valuable strength of our framework is that we can search the desired editing direction for only one frame and then edit the remaining frames with temporal consistency by moving the representative video identity feature. Additionally, the wild face video can be reliably edited by the advantage of the diffusion model that is superior in reconstruction to GAN. Finally, we have explored to optimize the semantic identity feature with CLIP loss for text-based video editing. Please refer to the Appendix D for the limitations and further discussion.

Acknowledgement This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST), No.2021-0-02068, Artificial Intelligence Innovation Hub, No.2022-0-00713, Meta-learning applicable to real-world problems, No.2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation) and the National Research Foundation of Korea (NRF) grants (No.2018R1A5A1059921) funded by the Korea government (MSIT). This work was also supported by KAIST-NAVER Hypercreative AI Center.

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 2

[2] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *arXiv preprint arXiv:2201.13433*, 2022. 1, 2, 3

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1, 3

[4] Cunjian Chen. Pytorch face landmark: A fast and accurate facial landmark detector, 2021. 5

[5] 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*, pages 4690–4699, 2019. 5

[6] Qianli Feng, Viraj Shah, Raghudeep Gadde, Pietro Perona, and Aleix Martinez. Near perfect gan inversion. *arXiv preprint arXiv:2202.11833*, 2022. 2

[7] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2, 3, 6, 11

[8] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 5

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 6

[11] 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. 1

[12] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, June 2022. 1, 3, 11, 12

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 11

[14] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 2

[15] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 13

[16] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 13

[17] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 3

[18] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 7, 11

[19] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 3

[20] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3, 7, 11

[21] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*, 2020. 5

[22] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1

[23] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 1, 2, 3, 4, 6, 13

[24] 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. 2

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3, 13

[26] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[27] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 2, 7, 13

[28] 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. 13

[29] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[30] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 1

[31] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, 2021. 3

[32] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[34] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2, 7

[35] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit H Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. *arXiv preprint arXiv:2201.08361*, 2022. 1, 2, 5, 6, 7, 11, 12, 13

[36] 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*, pages 10039–10049, 2021. 5

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[38] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 7

[39] Weihao Xia, Yujiu Yang, and Jing-Hao Xue. Gan inversion for consistent video interpolation and manipulation. *arXiv preprint arXiv:2208.11197*, 2022. 3

[40] Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally consistent semantic video editing. In *European Conference on Computer Vision*, pages 357–374. Springer, 2022. 3, 6, 7, 13

[41] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13789–13798, 2021. 1, 2, 6, 7, 12, 13

[42] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 6

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7