# Identity-Consistent Video De-identification via Diffusion Autoencoders

Yunhui Zhu[1*], Jingyi Cao[1*], Bo Liu[2], Tingxi Chen[1], Rong Xie[1,3], Li Song[1,3†]
[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] School of Computer Science, University of Technology Sydney, Sydney, Australia
[3] Cooperative Medianet Innovation Center (CMIC), Shanghai Jiao Tong University, Shanghai, China
{zhuyunhui, cjycaojingyi, ctx.ximi17, xierong, song_li}@sjtu.edu.cn, bo.liu@uts.edu.au
*Equal contribution. †Corresponding author.

***Abstract*—With the rise of deep learning and the widespread use of face recognition, face image privacy has become a critical research issue. Face de-identification is acknowledged as effective for protecting identity privacy. As media formats diversify, it is imperative to extend privacy protection to videos. Addressing the core problem of identity consistency between frames, we propose a video de-identification approach based on the diffusion model. We disentangle video features with a diffusion autoencoder, where the identity and motion features are encoded into high-level semantic spaces while background and other facial identity-independent features into low-dimensional random subcodes. The unified time-independent identity representation is used to achieve coherent video de-identification results. Compared to existing methods, our proposed approach demonstrates superior performance in terms of privacy protection effectiveness, identity consistency between frames, and utility.**

***Index Terms*—Diffusion Models, Feature Disentanglement, Face De-identification, Identity Protection**

## I. Introduction

The privacy protection of face images is an emerging research issue in computer vision, dedicated to addressing privacy infringements arising from the rise of deep learning technologies and the widespread application of face recognition. Face images are considered one of the most privacy-sensitive biometrics, closely linked to personal identity.

De-identification is effective in protecting the identity of face images and aims to make the original identity inaccessible to biometric recognition mechanisms by hiding, modifying, or removing personal identifiers. With the development of generative models, de-identification techniques based on deep learning have received a lot of attention.

Early methods including face restoration [1] or full-face generation [2] effectively remove sensitive information but struggle with maintaining visual similarity. To further enhance the utility, some of the research achieved the trade-off by adding additional constraints [3]. To achieve more refined identity modification, some methods [4], [5] proposed to disentangle the face representation and only modify the relatively independent identity without affecting other identity-independent attributes, which can better retain the visual similarity with the original image.

Previous research primarily focused on static face images, but as media formats diversify, identity protection technologies must adapt to a wider range of media formats. The primary challenge is to seamlessly modify the original video stream without inducing visual artifacts such as flickering or distortion between frames while maintaining temporal consistency. Common approaches include interpolating and smoothing between adjacent frames [6], learning mixed masks [7], and incorporating past frames into the input when predicting the current frame result [8]. Utilizing the redundancy and continuity of data in the video [9] facilitates the comprehensive reconstruction of the entire video stream, enhancing stability and continuity in video de-identification results.

In this work, we propose a video de-identification method based on diffusion model. A diffusion autoencoder is applied to disentangle video features, which maps identity and motion features into a high-level semantic space and encodes background and other identity-independent facial attributes into a low-dimensional space. We address the issue of inter-frame identity inconsistency by using a unified identity across the entire video.

Our contributions can be summarized as follows:

1) We propose a video de-identification approach capable of editing disentangled, time-independent identity, resulting in continuous and stable results.
2) Our work leverages the diffusion autoencoder to enhance the adaptability to a wider range of images with better generalization capabilities.
3) Our method outperforms baselines by generating high-quality de-identification results across diverse datasets, showcasing better inter-frame identity consistency.

## II. Related Works

### A. Diffusion Models

Denoising diffusion probabilistic models (DDPMs) [10] represent a significant advancement in the realm of generative models. The fundamental objective of DDPMs is to predict and associate the noise inherent in the input image. In contrast to GANs and the majority of Variational Autoencoders (VAEs) that encode input data into a low-dimensional space, DDPMs maintain a latent space with dimensions equal to those of the input. This unique characteristic contributes to achieving superior image fidelity and diversity, even though the generation process necessitates a substantial number of feed-forward steps. To solve the problem of slow sample generation

in DDPM, DDIM [11] defines the image sampling process as a non-Markov Chain to accelerate the generation process.

Prechakul et al. [12] proposed a Diffusion Autoencoder (DiffAE) that employs two forms of latent variables: one to represent useful high-level semantic information and another to represent residual low-level random information. Through semantic and random encoders, DiffAE achieves high-quality image reconstruction and provides advanced semantic representation for downstream tasks. Inspired by this, we aim to disentangle facial video features, extracting a unified identity representation and subsequently achieving high-quality face video de-identification.

### B. Face Video De-identification

In recent years, the trend of widespread use of face videos on social media has led to a new research domain focused on face de-identification. Existing approaches to this issue can be classified into two main types.

*1) Identity Swapping Techniques:* Substituting faces in videos for de-identification is a direct approach. This can involve using real or synthesized identities, with the latter offering more comprehensive privacy protection. Samarzija et al. [13] used pre-trained active appearance models to swap faces. Zhu et al. [14] used deepfake technology for medical video de-identification, swapping patients' faces with open-source characters. To enhance visual fidelity, more advanced identity-swapping methods have been proposed.

*2) Identity Disentanglement-Based Methods:* While prior methods achieved notable results, their reliance on auxiliary identities poses compliance challenges with strict regulations. Gross et al. [15] used a generative multi-factor model to factorize input images into identity and non-identity components. They applied a de-identification algorithm to the combined factorized data, reconstructing de-identified images using the model's bases. Ren et al. [16] employed a multi-task extension of GAN with a face anonymizer and a motion detector. IdentityMask [9] proposed to introduce the motion flow into video de-identification, which reconstructs video by motion and the de-identified first frame.

## III. Methods

We define video face de-identification as follows: Given a face video $V = \left(x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right)$, where $x^{(i)}$ denotes the $i$-th frame, the de-identification algorithm $\mathcal{F}$ modifies or conceals identity information in the original video, denoted as:

$$ID\left\{x^{(i)}\right\} \neq ID\left\{\mathcal{F}(x^{(i)})\right\} \quad 1 \leq i \leq n \tag{1}$$

Although each frame is processed individually during the de-identification process, to ensure coherence and stability in the generated video, the resulting de-identified video should exhibit temporal identity consistency. This implies that the identity between frames in the de-identified video should be maintained, formally expressed as:

$$ID(\mathcal{F}(x^{(i)})) = ID(\mathcal{F}(x^{(j)})) \quad \forall i, j \in [1, n] \tag{2}$$

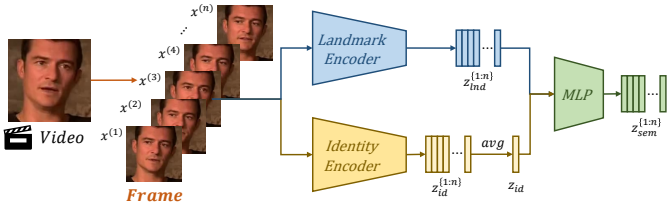


Fig. 1. The process of semantic features extraction for videos.

### A. Framework

We propose a framework for face video de-identification, leveraging diffusion models to learn a semantically meaningful latent space. De-identification is achieved by editing the disentangled identity features, preserving motion and other identity-independent attributes.

Initially, identity information $z_{\text{id}}$ and facial landmarks $z_{\text{lnd}}$ are extracted from the original image using the identity encoder $E_{\text{id}}$ and landmark encoder $E_{\text{lnd}}$. A mapping network, composed of a multilayer perceptron, yields a high-level semantic representation $z_{\text{sem}}$. The conditional DDIM functions dually as both an encoder and a decoder. During the encoding process, it transforms low-level random variations based on the semantic representation $z_{\text{sem}}$ into $x_T$, encompassing background features and other attributes unrelated to facial identity and motion. In the decoding process, it can generate the corresponding face image based on $z_{\text{sem}}$ and $x_T$.

### B. Video Feature Disentanglement

We refine the feature representation of $N$ frames of video denoted as $\left\{x_0^{(n)}\right\}_{n=1}^{N}$ into temporally invariant identity representation, temporally correlated motion representation, and frame-specific facial attributes and background representation.

Inspired by DiffAE [12], identity and motion-related information are suitable for mapping to a high-level semantic space, extracting relevant features $z_{\text{sem}}$. In contrast, frame-related background and identity-independent attribute details exhibit more diverse variations and are better suited for encoding into a low-level space $x_T$.

To achieve the aforementioned feature disentanglement, our approach includes two independent encoders: the identity encoder $E_{\text{id}}$ and the landmark encoder $E_{\text{lnd}}$. We utilize a mapping network $M$ to map them to a high-level semantic space, as illustrated in Fig. 1. Additionally, it incorporates a conditional noise predictor $\epsilon_\theta$ for the diffusion process.

The identity encoder extracts identity features for each frame $x_0^{(n)}$, and computes the mean of $N$ identity features as the unified identity for the entire video, as shown in Eq. (3).

$$z_{\text{id}}^{(n)} = E_{\text{id}}\left(x_0^{(n)}\right) \quad z_{\text{id}} = \frac{1}{N}\sum_{n=1}^{N} z_{\text{id}}^{(n)} \tag{3}$$

We utilize the landmark encoder to extract $z_{\text{lnd}}^{(n)}$ from each frame $x_0^{(n)}$ to characterize the motion, as formulated in Eq. (4).

$$z_{\text{lnd}}^{(n)} = E_{\text{lnd}}\left(x_0^{(n)}\right) \tag{4}$$

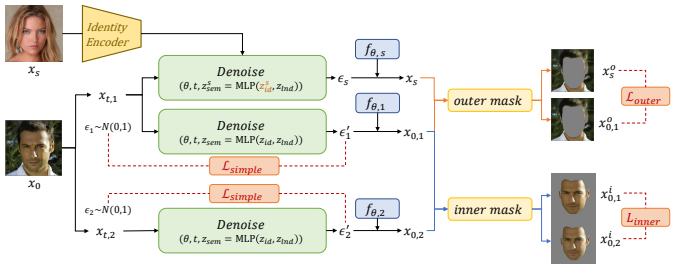


Fig. 2. Overview of the diffusion autoencoders training loss functions.

We use the pre-trained $E_{\text{id}}$ and $E_{\text{lnd}}$ to ensure the identity and motion disentangled from other facial features, and provide supervision for training. Subsequently, we use a learnable mapping network $M$ to map frame-independent $z_{\text{id}}$ and frame-specific $z_{\text{lnd}}$ to the semantic space, obtaining semantic features $z_{\text{sem}}^{(n)}$ corresponding to each frame, as formulated in Eq. (5).

$$z_{\text{sem}}^{(n)} = M(z_{\text{id}}, z_{\text{lnd}}^{(n)}) \tag{5}$$

Utilizing a noise estimator $\epsilon_\theta$ conditioned on $z_{\text{sem}}^{(n)}$, we encode other information contained in the face image, excluding semantic features, from the original image frames into random codes $x_T^{(n)}$. This process is conceptualized as a stochastic encoder $E_{\text{sto}}$, expressed as Eq. (6).

$$x_T^{(n)} = E_{\text{sto}}(x_0^{(n)}, z_{\text{sem}}^{(n)}) \tag{6}$$

It is important to note that the stochastic encoder $E_{\text{sto}}$ does not involve additional network structures. Instead, it relies on a deterministic forward process determined by conditional DDIM, as outlined in Eq. (7). The resulting noisy image $x_T^{(n)}$ maintains the same dimensions as the original image, facilitating the encoding of background information without sacrificing spatial details.

$$\begin{aligned} x_{t+1} &= \sqrt{\alpha_{t+1}} f_\theta\left(x_t, t, z_{\text{sem}}\right) + \sqrt{1-\alpha_{t+1}} \epsilon_\theta\left(x_t, t, z_{\text{sem}}\right) \\ f_\theta\left(x_t, t, z_{\text{sem}}\right) &= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\alpha_t}\epsilon_\theta\left(x_t, t, z_{\text{sem}}\right)\right) \end{aligned} \tag{7}$$

The deterministic execution of the reverse generation process through conditional DDIM, as Eq. (8), reconstructs the encoded features $\left(z_{\text{sem}}^{(n)}, x_T^{(n)}\right)$ back into the original frame. Through the semantic encoder and stochastic encoder, detailed representations of the original image are obtained, simultaneously providing advanced semantics for downstream tasks.

$$p_\theta\left(x_{0:T}^{(n)} \mid z_{\text{sem}}^{(n)}\right) = p\left(x_T^{(n)}\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1}^{(n)} \mid x_t^{(n)}, z_{\text{sem}}^{(n)}\right) \tag{8}$$

### C. Training Process

The training of the diffusion autoencoder is illustrated in Fig. 2, and the overall loss function consists of two components. The first part is the DDPM loss represented by Eq. (9).

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0,I), t} \left\| \epsilon_\theta\left(x_t^{(n)}, t, z_{\text{sem}}^{(n)}\right) - \epsilon_t \right\|_1, \tag{9}$$

where $z_{\text{sem}}^{(n)}$ represents the high-level semantic features of the input image $x_0^{(n)}$. We aim to encode relevant information of the facial region in the semantic features $z_{\text{sem}}^{(n)}$, with other background information encoded in $x_T$. Therefore, additional internal consistency loss $\mathcal{L}_{\text{inner}}$ and external consistency loss $\mathcal{L}_{\text{outer}}$ are designed to prevent leakage of facial identity information into $x_T$. During training, two different Gaussian noises, $\epsilon_1$ and $\epsilon_2$, are sampled for the original image $x_0$, resulting in two distinct noise samples $x_{t,1}$ and $x_{t,2}$. The objective is to minimize the difference between $f_{\theta,1}$ and $f_{\theta,2}$ in areas excluding the facial region, as formulated in Eq. (10):

$$\mathcal{L}_{\text{inner}} = \mathbb{E}_{x_0 \sim q(x_0)} \left\| f_{\theta,1} \odot m - f_{\theta,2} \odot m \right\|_1, \tag{10}$$

where $f_{\theta,i} = \frac{1}{\sqrt{\alpha_t}}\left(x_{t,i} - \sqrt{1-\alpha_t}\epsilon_\theta\left(x_{t,i}, t, z_{\text{sem}}\right)\right)$.

For the noise samples $x_t$, the denoising process is applied using the high-level semantic features $z_{\text{sem}}$ extracted from the original image $x_0$, resulting in a denoised image $\hat{x}_0 = f_{\theta,r}\left(\varepsilon_r\right)$. Additionally, the results $f_{\theta,s}$ are generated using identity information from another face image $x_s$, and by minimizing the difference in the background between $\hat{x}_0$ and $f_{\theta,s}$, as expressed in Eq. (11).

$$\mathcal{L}_{\text{outer}} = \mathbb{E}_{x_0 \sim q(x_0)} \left\| f_{\theta,r} \odot (1-m) - f_{\theta,s} \odot (1-m) \right\|_1, \tag{11}$$

where, $f_{\theta,r}$ represents the original reconstruction process and $f_{\theta,s}$ represents the generated result obtained by using the high-level semantic features $z_{\text{sem}}^s$ corresponding to the identity $z_{\text{id}}^s$ extracted from another face image $x_s$.

Combining the above loss functions, the total loss employed in our training can be expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \lambda_1 \mathcal{L}_{\text{inner}} + \lambda_2 \mathcal{L}_{\text{outer}}, \tag{12}$$

where $\lambda_1$ and $\lambda_2$ are hyperparameters.

## IV. Experiments

### A. Implement Details

*1) Network Architecture:* The overall model is based on an improved version of the DDIM [17] model, utilizing a U-Net structure composed of residual blocks and attention modules. The resolution of the model is set to 256×256. A pretrained face recognition model ArcFace [18] and a face keypoint detection model [1] are employed to obtain $z_{\text{id}} \in \mathcal{R}^{512}$ and $z_{\text{lnd}} \in \mathcal{R}^{106}$, respectively. These embeddings are further mapped to $z_{\text{sem}} \in \mathcal{R}^{512}$ through an MLP network. The temporal aspect during the diffusion process is initially embedded into a 128-dimensional vector using positional encoding, followed by projection into a 512-dimensional vector with a two-layer MLP network activated by SiLU.

*2) Dataset:* During training, we use the video dataset VoxCeleb [19], which comprises 22,496 videos obtained from YouTube. The video frames are aligned and cropped using the approach proposed by Tzaban et al. [20], resulting in videos ranging from 64 to 1024 frames with a resolution adjusted to 256×256.

[1] https://github.com/cunjian/pytorch_face_landmark

*3) Experiment Setup:* The approach we proposed is implemented using PyTorch and employs the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of $1 \times 10^{-4}$. The hyperparameters in Eq. (12) are set to $\lambda_1 = \lambda_2 = 1$. The entire model undergoes end-to-end training on a GeForce RTX 3090 GPU, with a batch size of 8.

### B. Qualitative Experiments

*1) Reconstruction Results:* We selected 10 videos featuring distinct identities from the RAVDESS dataset [21]. This dataset comprises 7,356 videos, including speaking videos with varied emotions and expressions from 24 professional actors. We employed two StyleGAN-based image inversion and reconstruction methods, Pixel2Style2Pixel (pSp) [22] and Encoder4Editing (e4e) [23] for baseline comparison, where the results are shown in Fig. 3. The results of StyleGAN-based methods exist minor changes in detail due to the accuracy of inverse latent codes and the influence of generative priors, and our approach can achieve more accurate results.

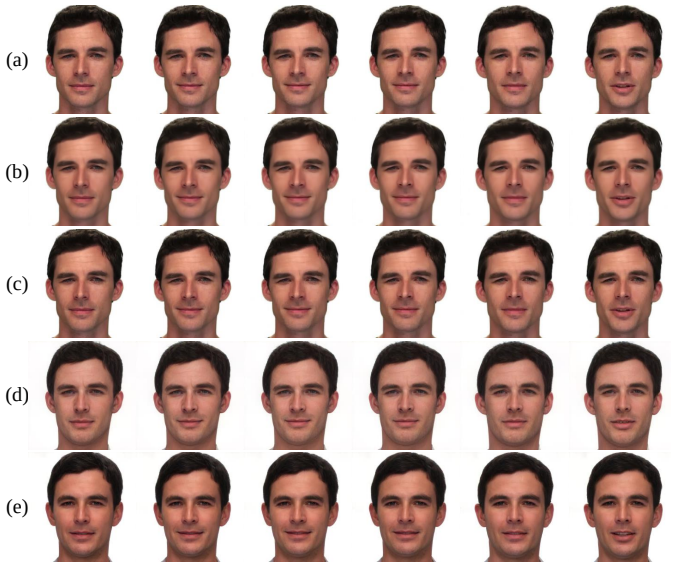


Fig. 3. Video reconstruction results and comparison with baseline models, where (a) the original video images (b) ours ($T = 100$) (c) ours ($T = 1000$) (d) Pixel2Style2Pixel [22] and (e) Encoder4Editing [23].

*2) De-identification Results:* During the de-identification process, our objective is to alter the identity while maintaining consistency in other information. In our framework, we selectively perturb the identity $z_{\text{id}}$, while preserving $z_{\text{lnd}}$ and other encoded information in $x_T$ unaltered. We first calculate the identity $z_{\text{id}}$, and then add random Gaussian noise to obtain the new identity $z'_{\text{id}} = z_{\text{id}} + n$. Subsequently, we generate de-identified results based on $z'_{\text{id}}$ and the original $x_T$.

We compared our method with others, including AMT-GAN [24], DeepPrivacy [2], CIAGAN [6], Gu et al. [25], and IdentityDP [4]. The results tested on the CelebA-HQ dataset prove that our method outperforms others in visual quality, demonstrating superior generalization and de-identification effectiveness while retaining more similarity.

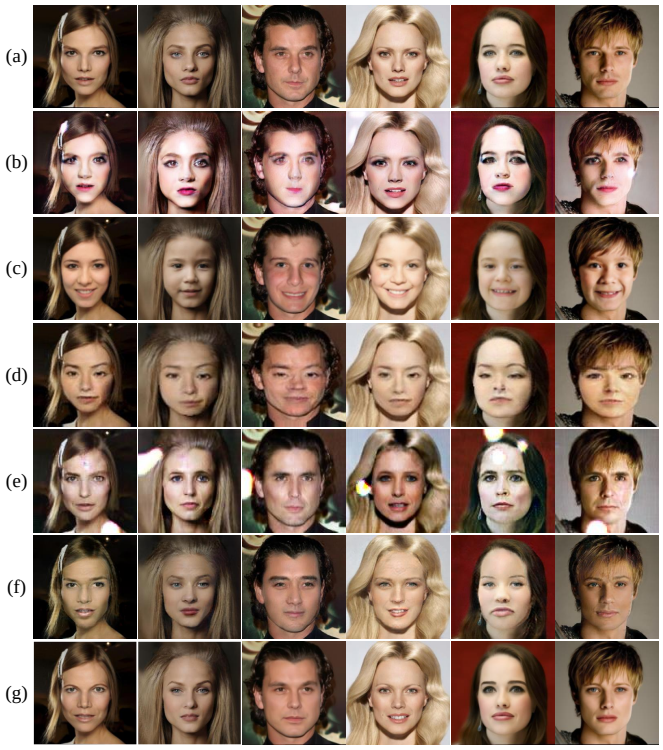


Fig. 4. Face image de-identification results and qualitative comparison of various state-of-the-art methods, where (a) original image (b) AMT-GAN [24] (c) DeepPrivacy [2] (d) CIAGAN [6] (e) Gu et al. [25] (f) IdentityDP [4] ($\varepsilon = 0.5$) (g) ours ($T = 100$).

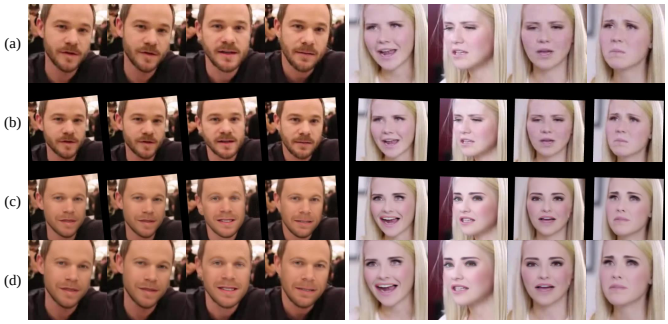


Fig. 5. The video de-identification results of our method, where (a) the original image (b) cropped and aligned result (c) the de-identified result, and (d) the inverse transform result.

The results of video de-identification are illustrated in Fig. 5. We initially perform image cropping and rotation alignment based on facial key points, and the de-identified faces are transformed back to the original frames through inverse transformations. The results prove that our approach can effectively handle video sequences, maintaining identity-independent attributes while successfully achieving identity feature manipulation. The de-identified video results exhibit consistent identity preservation across different frames.

*3) Ablation Experiments:* The ablation results for the loss functions in Eq. (12) are shown in Fig. 6. We mainly test the mask range of $\mathcal{L}_{\text{inner}}$ and whether applying $\mathcal{L}_{\text{outer}}$.

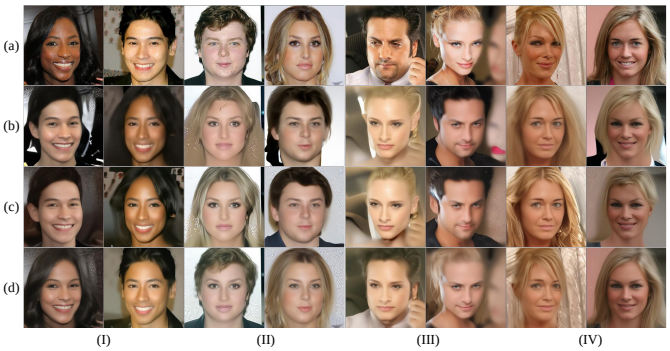


Fig. 6. The ablation results of our method with $T = 20$, where (a) the original image, (b)-(d) are the corresponding results of exchanging identity features in each group of images, (b) only use $\mathcal{L}_{\text{inner}}$ and the inner mask contains the full portrait foreground (c) only use $\mathcal{L}_{\text{inner}}$ and the inner mask contains only the range of faces (d) using both $\mathcal{L}_{\text{inner}}$ and $\mathcal{L}_{\text{outer}}$.

TABLE I
PERFORMANCE EVALUATION OF IMAGE RECONSTRUCTION AND COMPARISON WITH OTHER METHODS. THE RED REPRESENTS THE BEST.

| | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| Pixel2Style2Pixel [22] | 24.364 | 0.816 | 0.314 |
| Encoder4Editing [23] | 25.129 | 0.831 | 0.157 |
| Ours($T = 100$) | 34.909 | 0.948 | 0.246 |
| Ours($T = 1000$) | 40.676 | 0.994 | 0.065 |

Based on the experimental results, we can find that: (1) If the mask $m$ covers the entire foreground, it may result in black regions or hair part losses when the foreground in the target identity is larger than that of the source image, (2) better disentanglement can be obtained when simultaneously using both internal and external consistency losses, as swapping identity does not affect hairstyle or background while retaining facial features unrelated to identity, preserving higher visual similarity with the original image.

### C. Quantitative Experiments

*1) Reconstruction Results:* We randomly selected 500 face images from the CelebA-HQ dataset and 10 videos from the RAVDESS [21] for reconstruction testing. We compare the PSNR, SSIM and LPIPS to quantitatively assess reconstruction performance and the results are shown in Table I.

In comparison to GAN-based inversion methods, our framework achieves superior reconstruction results, preserving higher pixel-level similarity even with a relatively low number of sampling steps ($T = 100$). Increasing the sampling steps can further enhance the performance.

*2) De-identification Results:* We randomly selected 1,000 images from the CelebA-HQ dataset for testing. Face recognition models always determine whether two face images share the same identity information by computing the identity distance between their identity features and comparing it to a threshold, so we calculate the **identity distance** between the de-identification results with the original images to evaluate privacy protection effectiveness.

TABLE II
PRIVACY PROTECTION EFFECTIVENESS EVALUATION AND COMPARISON WITH OTHER METHODS WITH DIFFERENT FACE RECOGNITION MODELS. THE RED ONE REPRESENTS THE BEST AND THE BLUE ONE THE SECOND.

| *Id-distance* | ArcFace↓ | FR↑ | FaceNet↑ | |
|---|---|---|---|---|
| | | | VGGFace2 | CASIA |
| AMT-GAN [24] | 0.672 | 0.556 | 0.939 | 0.915 |
| DeepPrivacy [2] | 0.610 | 0.749 | 1.206 | 1.172 |
| CIAGAN [6] | 0.517 | 0.759 | 1.191 | 1.162 |
| Gu et al. [25] | 0.526 | 0.838 | 1.201 | 1.138 |
| IdentityDP [4] | 0.410 | 0.751 | 1.225 | 1.184 |
| Ours | 0.402 | 0.761 | 1.269 | 1.182 |

TABLE III
IDENTITY CONSISTENCY EVALUATION AND COMPARISON WITH METHODS WITH DIFFERENT FACE RECOGNITION MODELS. CONTRARY TO TABLE II, HIGHER IDENTITY SIMILARITY INDICATES BETTER CONSISTENCY.

| *Id-consistency* | ArcFace↑ | FR↓ | FaceNet↓ | |
|---|---|---|---|---|
| | | | VGGFace2 | CASIA |
| AMT-GAN [24] | 0.752 | 0.316 | 0.562 | 0.580 |
| DeepPrivacy [2] | 0.519 | 0.610 | 1.016 | 0.919 |
| CIAGAN [6] | 0.696 | 0.413 | 0.617 | 0.604 |
| Gu et al. [25] | 0.692 | 0.428 | 0.815 | 0.789 |
| IdentityDP [4] | 0.492 | 0.611 | 1.002 | 0.967 |
| Ours | 0.781 | 0.392 | 0.513 | 0.509 |

The results presented in Table II show that our method effectively achieves identity protection, demonstrating robustness across various face recognition models. Similar to IdentityDP [4], our algorithm adopts disentangled identity editing in latent space and exhibits more effective protection compared to adversarial perturbation methods [24] and full-face generation methods [2]. Compared to conditional GAN-based face de-identification methods [6], [25], our approach allows more targeted identity protection.

We selected 15 videos from VoxCeleb [19] and 15 from RAVDESS [21] for testing. For other baselines, we set the same conditions between each frame from the same video to maximize identity consistency. We choose the de-identification result of the first frame as a reference and calculate the identity distance between other de-identified frames of the same video as the measure of **identity consistency**, as shown in Table III. Compared with other methods, our method uses a uniform time-independent identity for each frame, which can better preserve identity consistency.

*3) Utility:* The utility evaluation results are presented in Table IV. We evaluate the utility for identity-independent computer vision tasks by performing face detection and defining the proportion of faces that can still be detected in the protected images as Face Detectability (FD). We further calculate the Pixel-level Distance (PD) of the face region between the de-identified results and the original image. We use peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) to measure image similarity at the pixel level and calculate LPIPS [26] to measure the perception similarity. It is evident that deep learning-based algorithms consistently achieve high-quality image generation, indicating a high detectability of faces. Due to the constraints in our design, the semantic

TABLE IV
THE UTILITY EVALUATION AND COMPARISON WITH OTHER METHODS UNDER DIFFERENT METRICS.

| | *FD* ↑ | *PD* ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|---|
| AMT-GAN [24] | 0.992 | 2.571 | 20.516 | 0.798 | 0.259 |
| DeepPrivacy [2] | 0.999 | 3.274 | 21.683 | 0.792 | 0.411 |
| CIAGAN [6] | 0.981 | 4.598 | 18.175 | 0.587 | 0.497 |
| Gu et al. [25] | 0.925 | 3.200 | 19.062 | 0.683 | 0.489 |
| IdentityDP [4] | 0.995 | 1.758 | 24.018 | 0.865 | 0.293 |
| Ours | 0.994 | 1.614 | 25.127 | 0.873 | 0.276 |

features correspond closely to the facial interior, resulting in de-identification results that can maintain a more consistent pixel-level distance within the facial region. Furthermore, the diffusion model's excellent generalization, achieved by accurately encoding background information in the random code $x_T$ within the disentanglement framework, contributes to higher pixel-level and perceptual similarity.

## V. CONCLUSION

In this paper, we present a face video de-identification framework, which addresses the challenge of identity coherence across frames in de-identified video. We encode disentangled facial identity and motion into the semantic space, while other information in video sequences is encoded into a low-dimensional space, facilitating high-quality image reconstruction. Inter-frame consistency is preserved by utilizing a consistent identity throughout the de-identified video generation. The diffusion model demonstrates superior generalization capabilities and effectively adapts to a broader range of image characteristics. Experimental results showcase that the proposed algorithm outperforms existing face de-identification methods. Future research may refine and extend the proposed framework for real-world applications, evaluating its performance in complex scenarios.

## ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Central Universities and MoE-China Mobile Research Fund Project (MCM20180702); in part by the 111 project under Grant B07022 and Sheitc No.150633; in part by the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

## REFERENCES

[1] Q. Sun, L. Ma, S. J. Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5050–5059.

[2] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *International Symposium on Visual Computing*. Springer, 2019, pp. 565–578.

[3] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-protective-gan for privacy preserving face de-identification," *Journal of Computer Science and Technology*, vol. 34, pp. 47–60, 2019.

[4] Y. Wen, B. Liu, M. Ding, R. Xie, and L. Song, "Identitydp: Differential private identification protection for face images," *Neurocomputing*, 2022.

[5] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Personalized and invertible face de-identification by disentangled identity information manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3334–3342.

[6] M. Maximov, I. Elezi, and L. Leal-Taixé, "Ciagan: Conditional identity anonymization generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5447–5456.

[7] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9378–9387.

[8] T. Balaji, P. Blies, G. Göri, R. Mitsch, M. Wasserer, and T. Schön, "Temporally coherent video anonymization through gan inpainting," *arXiv preprint arXiv:2106.02328*, 2021.

[9] Y. Wen, B. Liu, J. Cao, R. Xie, L. Song, and Z. Li, "Identitymask: Deep motion flow guided reversible face video de-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[11] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[12] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 609–10 619, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244729224

[13] B. Samarzija and S. Ribaric, "An approach to the de-identification of faces in different poses," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2014, pp. 1246–1251.

[14] B. Zhu, H. Fang, Y. Sui, and L. Li, "Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 414–420.

[15] R. Gross, L. Sweeney, J. F. Cohn, F. D. la Torre, and S. Baker, "Face de-identification," in *Protecting Privacy in Video Surveillance*, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:2043186

[16] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 620–636.

[17] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[18] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:8923541

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:10475843

[20] R. Tzaban, R. Mokady, R. Gal, A. H. Bermano, and D. Cohen-Or, "Stitch it in time: Gan-based facial editing of real videos," *SIGGRAPH Asia 2022 Conference Papers*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246063490

[21] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS ONE*, vol. 13, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:21704094

[22] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[23] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *arXiv preprint arXiv:2102.02766*, 2021.

[24] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 014–15 023.

[25] X. Gu, W. Luo, M. S. Ryoo, and Y. J. Lee, "Password-conditioned anonymization and deanonymization with face identity transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 727–743.

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.