DORMANT: Defending against Pose-driven Human Image Animation

Jiachen Zhou[1,2], Mingsi Wang[1,2], Tianlin Li[3], Guozhu Meng[1,2,*], Kai Chen[1,2,*]
[1]*Institute of Information Engineering, Chinese Academy of Sciences, China*
[2]*School of Cyber Security, University of Chinese Academy of Sciences, China*
[3]*Nanyang Technological University, Singapore*
{zhoujiachen, wangmingsi, mengguozhu, chenkai}@iie.ac.cn, tianlin001@e.ntu.edu.sg

Abstract

Pose-driven human image animation has achieved tremendous progress, enabling the generation of vivid and realistic human videos from just one single photo. However, it conversely exacerbates the risk of image misuse, as attackers may use one available image to create videos involving politics, violence, and other illegal content. To counter this threat, we propose DORMANT, a novel protection approach tailored to defend against pose-driven human image animation techniques. DORMANT applies protective perturbation to one human image, preserving the visual similarity to the original but resulting in poor-quality video generation. The protective perturbation is optimized to induce misextraction of appearance features from the image and create incoherence among the generated video frames. Our extensive evaluation across 8 animation methods and 4 datasets demonstrates the superiority of DORMANT over 6 baseline protection methods, leading to misaligned identities, visual distortions, noticeable artifacts, and inconsistent frames in the generated videos. Moreover, DORMANT shows effectiveness on 6 real-world commercial services, even with fully black-box access.

Warning: This paper contains unfiltered images generated by diffusion models that may be disturbing to some readers.

1 Introduction

Diffusion models such as Stable Diffusion [57], DALL·E [56], and Imagen [61] series models, have demonstrated their unprecedented capabilities in the field of image generation. More recently, video diffusion models like Stable Video Diffusion [7] and Sora [49], have also achieved significant advancements, capable of producing movie-quality and professional-grade videos. In particular, pose-driven human image animation methods have emerged [36], enabling the generation of controllable and realistic human videos by animating reference images according to desired pose sequences. The resulting videos maintain the appearance of the original reference

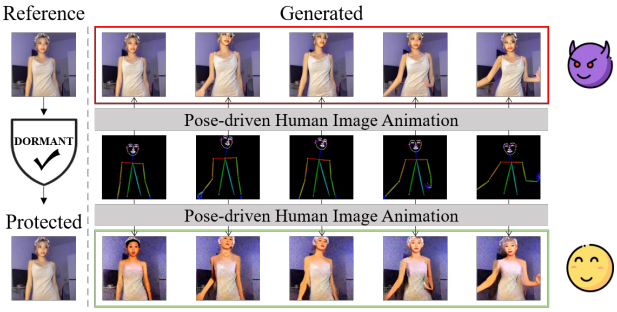


Figure 1: Illustration of defense against pose-driven human image animation. Generated video from the protected image displays mismatched identities and distorted backgrounds.

while accurately following the motion guidance provided by the poses. This innovation holds considerable potential across various applications, including social media, entertainment videos, the movie industry, and virtual characters, *etc.*

While pose-driven human image animation methods have revolutionized video generation, they also significantly lower the barriers to creating deceptive and malicious human videos, raising serious concerns about unauthorized image usage [15, 42, 52]. With just a single image of the victim, attackers can generate countless human videos, depicting the individual performing any action dictated by the pose input. For example, attackers can create dancing videos of celebrities to post on video platforms for commercial gains, fabricate fake videos of politicians to incite social harm, or produce Not Safe for Work (NSFW) videos that disrupt people's normal lives. These manipulated videos are easy and low-cost to generate, yet they severely violate portrait and privacy rights and potentially cause considerable harm to the unsuspecting individual.

Despite the severity of these hazards, there is a clear lack of research specifically aimed at preventing the misuse of pose-driven human image animation. To our knowledge, the most relevant concurrent work is VGMShield [50], which is also the only protection method designed to counteract malicious image-to-video generation. VGMShield applies protective

*Corresponding authors.

perturbation to the image, leading to low-quality video generation by producing incorrect or bizarre frames. The protective perturbation is optimized by targeting the encoding process of Stable Video Diffusion (SVD) [7], deceiving both the image and video encoders into misinterpreting the input image. However, this deviation in the embedding space is not sufficient to effectively disrupt the appearance features contained in images, thus failing to adequately protect human portraits. Furthermore, VGMShield demonstrates limited effectiveness and transferability when defending against non-SVD-based video generation methods, including various pose-driven human image animation techniques.

While few studies focus specifically on the misuse of image-to-video generation, several protection methods exist to defend against text/image-to-image generation [41, 62, 63, 68, 75, 84]. These methods typically optimize protective perturbations by targeting the fine-tuning process of customization techniques [59], the encoding process of the Variational AutoEncoder [19], or the reverse process of the denoising UNet [26]. However, current protections against image generation are inadequate for defending against pose-driven human image animation, as they fail to effectively deceive the advanced feature extractors used in animation methods to disrupt appearance features and break the temporal consistency across frames that is essential in video generation.

In this paper, we propose DORMANT (Defending against Pose-driven Human Image Animation), to address this notable gap in effective protection methods that prevent unauthorized image usage in pose-driven human image animation, thereby safeguarding portrait and privacy rights. As illustrated in Figure 1, DORMANT applies protective perturbation to the reference image, resulting in the protected image that appears visually similar to the original but leads to poor-quality video generation when used, causing issues including mismatched appearances, distorted visuals, noticeable artifacts, and incoherent frames. We assume that black-box access to the animation methods and models that the potential attacker might use, and employ a transfer-based strategy to optimize the protective perturbation according to our proposed objective function, utilizing publicly available pre-trained models as surrogates. The optimization objective includes: 1) feature misextraction, which induces deviations in the latent representation and comprehensively disrupts the semantic and fine-grained appearance features of the reference image; and 2) frame incoherence, which targets the appearance alignment and temporal consistency among the generated video frames. We further adopt Learned Perceptual Image Patch Similarity (LPIPS) [79], Expectation over Transformation (EoT) [5], and Momentum [18] to enhance the imperceptibility, robustness, and transferability of the protective perturbation, respectively.

We evaluate the protection performance of DORMANT on 8 cutting-edge pose-driven human image animation methods. The experimental results across 6 image-level and video-level generative metrics demonstrate the superior effectiveness of DORMANT, compared to 6 baseline protections against image and video generation. We conduct experiments on 4 datasets, covering tasks such as human dance generation, fashion video synthesis, and speech video generation. Results of human and GPT-4o [48] studies further highlight the superiority of DORMANT from the perspectives of human perception and multimodal large language model. DORMANT also exhibits remarkable robustness against 5 popular transformations and 6 advanced purifications. To further assess the transferability of DORMANT, we conduct additional experiments on 4 image-to-image techniques, 4 image-to-video techniques, and 6 real-world commercial services, DORMANT effectively defends against these various generative methods.

Contributions. We make the following contributions:

- We address the challenge of effectively preventing unauthorized image usage in pose-driven human image animation, safeguarding individuals' rights to portrait and privacy.
- We design novel objective functions to optimize the protective perturbation, inducing misextraction of appearance features and incoherence among generated video frames.
- Extensive experiments conducted on 8 pose-driven human image animation methods, 4 image-to-image methods, 4 image-to-video methods, and 6 commercial services highlight the effectiveness and transferability of DORMANT.

2 Background

2.1 Latent Diffusion Model

Different from the Denoising Diffusion Probabilistic Model (DDPM) [26] which directly operates in pixel space, to reduce the computational demand of Diffusion Model (DM), the Latent Diffusion Model (LDM) [57] applies DM training and sampling in the latent space, thus striking a balance between image quality and throughout. LDM utilizes the Variational AutoEncoder (VAE) [19] to provide the low-dimensional latent space, which consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. More precisely, given an image x, the encoder $\mathcal{E}$ maps it to a latent representation: $z = \mathcal{E}(x)$, and then the decoder $\mathcal{D}$ reconstructs it back to the pixel: $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$.

LDM performs the diffusion process proposed in DDPM within the latent space, where the forward process iteratively adds Gaussian noise to the latent representation to make it noisy, while the reverse process predicts and denoises applied noise using a denoising UNet [58]. Specifically, given an image x, during the forward process, its latent representation $z_0 = \mathcal{E}(x)$ is perturbed with Gaussian noise over T timesteps, transforming z_0 into a standard Gaussian noise z_T:

$$
\begin{aligned}
q(z_{1:T}|z_0) &= \prod_{t=1}^{T} q(z_t|z_{t-1}), \\
q(z_t|z_{t-1}) &= \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \boldsymbol{I}),
\end{aligned}
\tag{1}
$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is the variance schedule. By defining $\alpha_t = 1-\beta_t$ and $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, we can use the reparameterization technique [34] to sample z_t at timestep t as follows:

$$\begin{aligned} q(z_t|z_0) &= \mathcal{N}(z_t; \sqrt{\overline{\alpha}_t} z_0, (1-\overline{\alpha}_t)\boldsymbol{I}), \\ z_t &= \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1-\overline{\alpha}_t}\,\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \boldsymbol{I}). \end{aligned} \tag{2}$$

The reverse process aims to denoise the noisy latent representation z_T back to its original state with the following transition starting at $p(z_T) = \mathcal{N}(z_T; 0; \boldsymbol{I})$:

$$\begin{aligned} p_\theta(z_{0:T}) &= p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1}|z_t), \\ p_\theta(z_{t-1}|z_t) &= \mathcal{N}(z_{t-1}; \mu_\theta(z_t, c, t), \Sigma_\theta(z_t, c, t)), \end{aligned} \tag{3}$$

where c represents the embedding of conditional information such as text embedding obtained from CLIP ViT-L/14 text encoder [54], the mean $\mu_\theta(z_t, c, t)$ and the variance $\Sigma_\theta(z_t, c, t)$ are computed by the denoising UNet ε_θ parameterized by θ. With Eq. (2), Ho *et al.* [26] simplify the learning objective of the denoising UNet ε_θ in the ε-prediction form, where ε_θ is trained to predict the added noise ε_t at timestep t:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0, c, \varepsilon_t \sim \mathcal{N}(0,1), t} \Big[\|\varepsilon_t - \varepsilon_\theta(z_t, c, t)\|_2^2 \Big]. \tag{4}$$

Once trained, given $\tilde{z}_T$ sampled from the Gaussian distribution, the denoising UNet ε_θ can be utilized to predict ε_T and computed $\tilde{z}_{T-1}$ according to Eq. (3). Subsequently, $\tilde{z}_{T-2}, \tilde{z}_{T-3}, \ldots, \tilde{z}_1, \tilde{z}_0$ are progressively calculated, and finally $\tilde{z}_0$ is decoded by $\mathcal{D}$ to generate the image $\tilde{x} = \mathcal{D}(\tilde{z}_0)$.

2.2 LDM for Human Image Animation

Pose-driven human image animation [30, 36, 69, 71, 74, 86] is an image-to-video task aimed at generating a human video by animating a static human image based on a user-defined pose sequence. As shown in Figure 1, the created video needs to accurately align the appearance details from the reference image while following the motion guidance from the pose sequence, which can be produced by DWPose [76] or DensePose [22]. Recently, LDM-based animation methods have demonstrated superior effectiveness in addressing challenges in appearance alignment and pose guidance, enabling the generation of realistic and coherent videos.

Using LDM as the backbone, Stable Diffusion (SD) has achieved remarkable success in text/image-to-image generation tasks [57]. Pre-trained SD models effectively capture high-quality content priors. However, their network architecture is inherently designed for image generation, lacking the capability to handle the temporal dimension required for video generation. To extend foundational text/image-to-image models for image animation, many works [13, 30, 70, 74, 86] have incorporated temporal layers [23] into the denoising UNet for temporal modeling. Specifically, the feature map $X \in \mathbb{R}^{b \times f \times h \times w \times c}$ is reshaped to $X \in \mathbb{R}^{(b \times h \times w) \times f \times c}$, after which temporal attention is performed to capture the temporal dependencies among frames. Recently, some studies [53, 69, 82] have also adopted Stable Video Diffusion [7] as the backbone for image animation. SVD is an open-source image-to-video LDM trained on a large-scale video dataset; its strong motion-aware prior knowledge facilitates maintaining temporal consistency in image animation.

While pose-driven human image animation has demonstrated impressive capabilities in creating vivid and realistic human videos, it also raises serious concerns about unauthorized image usage by malicious attackers. With just a single photo of the victim, easily obtained from the web or social media, the malicious attacker can generate an unlimited number of human videos, manipulating the victim to perform any poses and thereby severely violating the individual's rights to publicity and privacy. These manipulated videos could be used for commercial or political purposes, potentially inflicting significant harm on the unsuspecting victim.

2.3 Protection Methods against LDM

While few studies addressing the issue of unauthorized image-to-video generation, prior research has proposed various protection methods [41, 62, 63, 68, 75, 84] against LDM to prevent image misuse in text/image-to-image generation. In the context of image generation, there are two primary LDM-based image misuse scenarios: concept customization and image manipulation. Existing protection methods mainly utilize Projected Gradient Descent (PGD) [45] to generate adversarial examples for LDM, thereby safeguarding against malicious image generation in both misuse scenarios. The defender aims to introduce protective perturbation δ to the image x, such that the protected image $x_p = x + \delta$ can deceive the LDM, resulting in poor-quality generation.

Defending against Concept Customization. Concept customization techniques [29, 59] fine-tune a few parameters of a pre-trained text-to-image model so that it quickly acquires a new concept given only 3-5 images as reference. The training loss of the most popular work DreamBooth [59] introduces a prior preservation loss to $\mathcal{L}_{LDM}$ in Eq. (4):

$$\begin{aligned} \mathcal{L}_{DB} = \mathbb{E}_{z_0, c, \varepsilon_t, \varepsilon'_{t'}, t, t'} \Big[&\|\varepsilon_t - \varepsilon_\theta(z_{t+1}, c, t)\|_2^2 \\ &+ \lambda \left\| \varepsilon'_{t'} - \varepsilon_\theta(z'_{t'+1}, c_{pr}, t') \right\|_2^2 \Big], \end{aligned} \tag{5}$$

where $z'_{t'+1}$ is the noisy latent representation of the class sample x' generated from LDM with prior prompt c_{pr}, λ controls for the weight of the prior term, which prevents over-fitting and text-shifting problems. Depending on the conceptual properties, there are two variants: style mimicry [63] and subject recontextualization [3]. Style mimicry seeks to generate paintings in an artist's style without consent, while subject recontextualization aims to memorize a subject (*e.g.*, a portrait

of a victim) and generate it in a new context. Protection methods [68, 84] mainly disrupt the learning process in Eq. (5) by solving the bi-level optimization: $\delta = arg\,max_{\delta}\mathcal{L}_{LDM}(\mathcal{E}(x+\delta),\theta'), s.t. \theta' = arg\,min_{\theta}\mathcal{L}_{DB}(\mathcal{E}(x+\delta),\theta)$, aimed at finding protective perturbation δ which degrades the personalization generation ability of DreamBooth, thereby preventing unauthorized concept customization.

Defending against Image Manipulation. Image Manipulation techniques [8, 9, 16, 46] leverage pre-trained image-to-image models to directly manipulate a single image, altering its appearance or transferring its style. Protection methods aimed at defending against image manipulation focus on disrupting the encoding process ($x \rightarrow z_0$) [62, 63] and the reverse process ($z_T \rightarrow z_0$) [41, 75]. The protective perturbation δ is optimized by fooling the VAE $\delta = arg\,max_{\delta}\|\mathcal{E}(x+\delta)-\mathcal{E}(x)\|_2^2$ and the denoising UNet $\delta = arg\,max_{\delta}\mathcal{L}_{LDM}(\mathcal{E}(x+\delta))$, thus making the LDM to generate images significantly deviating from original image x.

There are several differences between concept customization, image manipulation, and pose-driven human image animation. First, concept customization and image manipulation are image generation tasks, whereas human image animation is categorized as image-to-video generation. Second, concept customization requires multiple images (typically 3-5), while the other two techniques operate on a single image. Third, concept customization involves fine-tuning the LDM to memorize and generate a new concept. In contrast, image manipulation and human image animation leverage frozen pre-trained LDM directly for image or video generation without additional learning. Fourth, concept customization uses images as training data for fine-tuning, image manipulation directly edits the given image itself, while human image animation extracts appearance features from the reference image to guide video generation, rather than using the image directly.

These various ways of using images necessitate the design of specific objective functions for each case when optimizing the protective perturbation. Existing protections against concept customization and image manipulation mainly target the fine-tuning process, or the encoding and reverse processes, respectively. However, these methods are ineffective at disrupting the appearance features contained in human images. As shown in Figure 3 in Section 4.2, only DORMANT effectively induces noticeable mismatches in identity appearance between reference images and generated videos. The reason is that protective perturbations optimized with current objectives have a limited protective effect on various well-trained feature extractors used in animation methods. The ablation study results in Figure 11 in Section 4.6 also emphasize the need for specifically designed objective functions to induce feature misextraction, which existing methods lack. Moreover, they also neglect to break the temporal consistency across frames specifically required in video generation. As a result, current protections against image generation are insufficient for defending against pose-driven human image animation.

While many protection methods exist to prevent unauthorized image usage in text/image-to-image generation, there is a notable gap in defending against image-to-video generation. To our knowledge, the only concurrent work addressing this issue is VGMShield [50], which aims to prevent malicious image-to-video generation using SVD [7] by generating adversarial examples. More precisely, it attacks the encoding process of SVD, aiming at deceiving the image encoder and video encoder into misinterpreting the image. However, this deviation specifically in the embedding space of SVD is insufficient to effectively disrupt the appearance features in human images and shows limited effectiveness and transferability when defending against non-SVD-based video generation methods, including various pose-driven human image animation techniques. Overall, the ineffectiveness of existing protections against image and video generation highlights the need for novel protection methods specifically designed to prevent malicious pose-driven human image animation.

3 Methodology

DORMANT prevents image misuse in pose-driven human image animation by adding protective perturbation δ to the human image x, resulting in the protected image $x_p = x+\delta$. x_p is visually similar to x but would lead to poor-quality video generation if used, causing issues such as mismatched appearances, visual distortions, and frame inconsistencies. We assume that a fully black-box setting for the animation methods and models that potential attackers might use and employ a transfer-based strategy to optimize δ according to our proposed objective function $\mathcal{L}_{\text{DORMANT}}$, relying on some surrogate pre-trained models (*e.g.*, the publicly available CLIP model [54]) to which the defender has white-box access. As shown in Figure 2, $\mathcal{L}_{\text{DORMANT}}$ contains: 1) $\mathcal{L}_{vae}$ and $\mathcal{L}_{feature}$, which aim to induce misextraction of appearance features from the reference image; and 2) $\mathcal{L}_{frame}$, which is designed to cause incoherence among the generated video frames. Additionally, we adopt LPIPS [79], EoT [5], and Momentum [18], to further enhance the imperceptibility, robustness, and transferability of δ, respectively. The protective perturbation is optimized using PGD [45], which iteratively updates δ with a step size γ under an L_∞ norm constraint η as follows:

$$\delta_i = \delta_{i-1} + \gamma \cdot \text{sign}(\nabla_{\delta}\mathcal{L}_{\text{DORMANT}}),\ s.t.\ \|\delta\|_\infty \leq \eta. \quad (6)$$

3.1 Threat Model

Attacker. The attacker aims to generate human videos using pose sequences based on images obtained without the owner's consent. These fabricated videos could be misused for commercial or political purposes, violating the victim's rights to portrait and privacy. We assume that the attacker maliciously acquires multiple images of the victim from the web or social media. The attacker is aware of potential protections and can

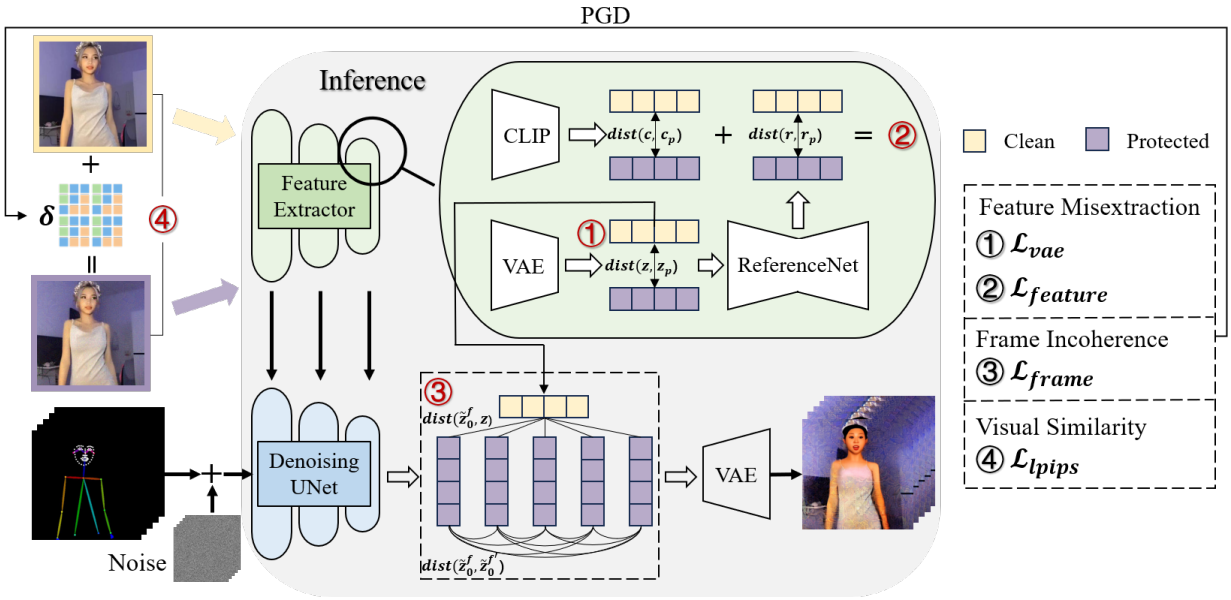


Figure 2: Overview of DORMANT. Here, we present the four components of our proposed objective function $\mathcal{L}_{\text{DORMANT}}$, which includes: $\mathcal{L}_{vae}$ and $\mathcal{L}_{feature}$ for feature misextraction, $\mathcal{L}_{frame}$ for frame incoherence, and $\mathcal{L}_{lpips}$ for visual similarity.

apply various countermeasures to these images to craft a suitable reference image for video generation. The attacker also has access to various LDMs for pose-driven human image animation, and significant computational power to run them.

Defender. We assume that the defender intends to apply protective perturbation to a human image before posting it online or on social media, to prevent potential unauthorized human video generation and thereby safeguard portrait and privacy rights. The protected image appears visually similar to the original image to human perception. However, any videos generated using this protected image would display mismatches in identity appearance and degraded quality, such as distorted backgrounds, visible artifacts, and incoherent frames, *etc*. To optimize the protective perturbation, the defender has access to some publicly available feature extractors and LDMs for pose-driven human image animation. However, the defender is unaware of, and therefore unable to approximate, the specific animation methods and models that the attacker might use. The defender could be either the image owner or a trustworthy third party with sufficient computing resources to execute the protection method effectively.

3.2 Feature Misextraction

As mentioned in Section 2.3, the mainstream usage of the reference image in pose-driven human image animation involves extracting its appearance features to guide video generation. In this context, the objective of optimizing the protective perturbation δ is to disrupt these appearance features, causing the feature extractor to extract incorrect features from the protected image. As a result, videos generated with this faulty guidance would differ from the reference image, thereby preventing unauthorized usage in human video generation.

VAE Encoder. As mentioned in Section 2.1, to reduce computational demands, LDM first maps the input image to a latent representation using the VAE encoder before performing the diffusion process. Given this, an intuitive strategy for inducing feature misextraction is to disrupt the encoding process, causing the encoder $\mathcal{E}$ to map the protected image x_p to a "wrong" latent vector $z_p = \mathcal{E}(x+\delta)$ that deviates significantly from the original latent $z = \mathcal{E}(x)$. Consequently, this misencoded latent z_p would contain incorrect appearance features, thus misleading the video generation process. The optimization objective $\mathcal{L}_{vae}$ is to maximize the distance between x and x_p in the VAE latent space, as defined by:

$$\underset{\|\delta\|_\infty \leq \eta}{arg\,max} \mathcal{L}_{vae} = \underset{\|\delta\|_\infty \leq \eta}{arg\,max} \|\mathcal{E}(x+\delta) - \mathcal{E}(x)\|_2^2. \quad (7)$$

We utilize the encoder $\mathcal{E}$ from *sd-vae-ft-mse* [1], a VAE fine-tuned on an enriched dataset with images of humans to improve the reconstruction of faces.

However, the deviation of the latent representation alone is insufficient to induce a deep-level misextraction of the appearance features. This is because pose-driven human image animation methods typically have stringent requirements for appearance alignment and often use pre-trained or custom-designed models for additional feature extraction. As a result, the attack on the VAE encoder has only a limited effect on these advanced feature extractors. To comprehensively disrupt the appearance features of the image and thus improve the transferability of the protective perturbation against various unknown feature extractors, we further employ CLIP image

encoder for semantic feature extraction and ReferenceNet for fine-grained feature extraction during the optimization of δ.

CLIP Image Encoder. The CLIP model [54] is trained on a diverse set of text-image pairs, and its image encoder $\mathcal{C}$ can be utilized to extract semantic features from the reference image x, which can then be integrated into the Transformer Blocks of the denoising UNet to guide video generation through cross-attention [70]. To ensure that the protected image $x_p = x + \delta$ contains different semantic features from the original image x, we further employ the CLIP image encoder $\mathcal{C}$ from *sd-image-variations* [35] for feature extraction during the optimization of δ. The objective is to maximize the distance between x and x_p in the CLIP embedding space, defined as follows:

$$\underset{\|\delta\|_\infty \leq \eta}{arg\,max} \|\mathcal{C}(x+\delta) - \mathcal{C}(x)\|_2^2. \tag{8}$$

However, CLIP image encoder is known to be less effective at capturing fine-grained details [30, 74]. There are two main factors for this limitation. First, CLIP is trained to match semantic features between text and images, which are typically sparse and high-level, resulting in a deficit of detailed features. Second, the CLIP image encoder only accepts low-resolution (224×224) images as inputs, leading to a loss of fine-grained information. Therefore, relying solely on the CLIP image encoder for optimizing δ is insufficient, as it mainly extracts semantic features while missing fine-grained details.

ReferenceNet. Inspired by the recent success of ReferenceNet in detailed feature extraction [13, 30, 66, 69, 74, 86], we further utilize ReferenceNet $\mathcal{R}$ as the feature extractor to optimize δ. ReferenceNet is a copy of the denoising UNet from SD used in image generation[1], designed to extract appearance features particularly low-level details from the reference human image. Specifically, the feature map $X_1 \in \mathbb{R}^{h \times w \times c}$ from the ReferenceNet is repeated by f times and concatenated with the feature map $X_2 \in \mathbb{R}^{f \times h \times w \times c}$ from the denoising UNet along w dimension. Then spatial attention is performed to transmit the appearance information from the reference image spatially. ReferenceNet inherits weights from the original SD and is further trained on human video frames, enabling it to provide fine-grained and detailed features essential for preserving the reference appearance in video generation.

In addition to using CLIP image encoder $\mathcal{C}$ to extract semantic features and optimize the protective perturbation δ according to Eq. (8), we also incorporate ReferenceNet $\mathcal{R}$ into the optimization process to enhance fine-grained feature extraction. To ensure that the protected image x_p exhibits distinct detailed features from the perspective of ReferenceNet, we maximize the distance between the corresponding extracted features $\mathcal{R}(\mathcal{E}(x+\delta))$ and $\mathcal{R}(\mathcal{E}(x))$. Moreover, to further improve the transferability of the protective perturbation, we optimize δ using an ensemble [18, 43] of $K = 3$ different pre-trained ReferenceNets [13, 30, 74]. The objective $\mathcal{L}_{feature}$ for feature misextraction using the CLIP image encoder and ReferenceNets is thus defined as follows:

$$\begin{aligned} \underset{\|\delta\|_\infty \leq \eta}{arg\,max} \mathcal{L}_{feature} = \underset{\|\delta\|_\infty \leq \eta}{arg\,max} &\|\mathcal{C}(x+\delta) - \mathcal{C}(x)\|_2^2 \\ &+ \sum_{k=1}^{K} \|\mathcal{R}_k(\mathcal{E}(x+\delta)) - \mathcal{R}_k(\mathcal{E}(x))\|_2^2. \end{aligned} \tag{9}$$

3.3 Frame Incoherence

While feature misextraction mainly focuses on disrupting the alignment of reference appearance, it is also crucial to consider the requirement of maintaining temporal consistency across frames in video generation when designing the optimization objective. Based on this, we further propose $\mathcal{L}_{frame}$, which attacks the video generation process of the denoising UNet to induce incoherence among the generated video frames, thus enhancing the protective effect. This requires the defender to provide both the reference image and the pose sequence to simulate the attacker's animation process. However, predicting the exact pose sequences the attacker might use is challenging. To this end, we directly extract the corresponding pose from the reference image and repeat it F times as the pose sequence to guide video generation.

The protective perturbation δ can be optimized from two perspectives: 1) maximizing the distance between each video frame and reference image to disrupt appearance alignment; and 2) maximizing the distance between each video frame and others to disrupt video consistency. Since pose-driven human image animation operates in the latent space, these two objectives can be formulated as increasing the distance between the latent vector of the reference image $z = \mathcal{E}(x)$ and that of each generated video frame $\tilde{z}_0^f$, or the distance between the latent vectors of different frames $\tilde{z}_0^f$ and $\tilde{z}_0^{f'}$. The objective $\mathcal{L}_{frame}$ for frame incoherence is defined as follows:

$$\begin{aligned} \underset{\|\delta\|_\infty \leq \eta}{arg\,max} \mathcal{L}_{frame} = \underset{\|\delta\|_\infty \leq \eta}{arg\,max} &\frac{1}{F} \sum_{f=1}^{F} \left\| \tilde{z}_0^f - \mathcal{E}(x) \right\|_2^2 \\ &+ \frac{2}{F(F-1)} \sum_{f=1}^{F} \sum_{f'=f+1}^{F} \left\| \tilde{z}_0^f - \tilde{z}_0^{f'} \right\|_2^2. \end{aligned} \tag{10}$$

We set $F = 5$ by default. As mentioned in Section 2.1, given the total inference steps T, we can sample $\tilde{z}_T$ from the Gaussian distribution and compute $\tilde{z}_0$ progressively from timestep T to timestep 0 according to Eq. (3). However, this would result in substantial GPU memory usage and increased time costs. To improve efficiency, we directly estimate $\tilde{z}_0$ using the reparameterization technique outlined in Eq. (2), rather than computing it step by step. Specifically, given a timestep t, $\tilde{z}_0$ is estimated by $\tilde{z}_0 = (\tilde{z}_t - \sqrt{1-\overline{\alpha}_t}\epsilon_t)/\sqrt{\overline{\alpha}_t}$, where $\overline{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is the variance schedule, ϵ_t is predicted by a pre-trained denoising UNet[2],

[1] ReferenceNet is identical to traditional denoising UNet used in image generation, while denoising UNet for human image animation incorporates additional temporal layers to generate video, as mentioned in Section 2.2.

[2] https://github.com/MooreThreads/Moore-AnimateAnyone

and $\tilde{z}_t$ is sampled from the Gaussian distribution. In each PGD iteration, we randomly sample one timestep t from the last 10 timesteps of T and estimate the corresponding $\tilde{z}_0$.

3.4 DORMANT

Overall, to optimize the protective perturbation δ for preventing unauthorized image usage in pose-driven human image animation, we propose $\mathcal{L}_{vae}$ and $\mathcal{L}_{feature}$ for feature misextraction, $\mathcal{L}_{frame}$ for frame incoherence. We also apply several techniques to further enhance the perturbation δ during optimization, including LPIPS [79] for imperceptibility, EoT [5] for robustness, and Momentum [18] for transferability.

LPIPS. In addition to constraining the protective perturbation δ using the L_∞ norm, we also incorporate LPIPS to further improve the imperceptibility of δ. LPIPS utilizes deep features from a pre-trained network to measure image distance, aligning more closely with human perception. Thus we use LPIPS to ensure that the protected image x_p remains visually similar to the original image x. Specifically, we include the following regularization term and minimize $\mathcal{L}_{lpips}$ during optimization:

$$\mathcal{L}_{lpips} = \max(\text{LPIPS}(x+\delta, x) - \zeta, 0), \tag{11}$$

where ζ denotes the budget of $\mathcal{L}_{lpips}$, and we set $\zeta = 0.1$ by default. During the optimization of δ, if the perceptual distance between x_p and x exceeds the bound ζ, $\mathcal{L}_{lpips}$ introduces a penalty to maintain visual similarity; otherwise, $\mathcal{L}_{lpips} = 0$.

EoT. To enhance the robustness of δ against various transformations, we adopt EoT into the PGD process. EoT introduces a distribution of transformation functions to the input and optimizes the expectation of the objective function values over these transformations. To reduce time and computational costs, during each PGD step, we randomly sample one of the following transformations and apply it to the protected image x_p: Gaussian blur, JPEG compression, Gaussian noise, Random resize (resizing to a random resolution and then back to the original size), or no transformation. These selected transformations are commonly adopted as potential countermeasures in prior studies and have been shown to weaken the effect of protective perturbations [3, 27, 83].

Momentum. Momentum enhances the transferability of the perturbation by accumulating a velocity vector in the gradient direction of the loss function across iterations. The memorization of previous gradients helps to stabilize updated directions and escape from poor local maxima. We incorporate momentum into the PGD process in Eq. (6) as follows:

$$\begin{aligned} g_i &= \mu \cdot g_{i-1} + \frac{\nabla_\delta \mathcal{L}_{\text{DORMANT}}}{\text{mean}(\|\nabla_\delta \mathcal{L}_{\text{DORMANT}}\|)}, \\ \delta_i &= \delta_{i-1} + \gamma \cdot \text{sign}(g_i), \end{aligned} \tag{12}$$

where g_i denotes the accumulated velocity vector at PGD step i, and μ is the decay factor controlling the impact of previous gradients, set to 0.5 by default. In each iteration, the current gradient is normalized using the mean of its absolute values rather than the L_1 distance used in the original paper [18].

Algorithm 1 DORMANT

Input: A human image x, PGD iterations N, step size γ, budget η, decay factor μ; pre-trained models to calculate $\mathcal{L}_{\text{DORMANT}}$: VAE encoder $\mathcal{E}$, CLIP image encoder $\mathcal{C}$, ensemble of ReferenceNets $\mathcal{R}_{1:K}$, and denoising UNet ϵ_θ

Output: The protected image x_p

Initialize $g_0 \leftarrow 0$ and $\delta_0 \leftarrow uniform(-\eta, \eta)$
for $i = 1$ to N **do**
  Sample transformation $\mathcal{T}$
  Sample timestep t
  Sample $\tilde{z}_t^{1:F} \sim \mathcal{N}(0, \boldsymbol{I})$
  Calculate $\mathcal{L}_{\text{DORMANT}}$ by Eq. (13)
  $g_i \leftarrow \mu \cdot g_{i-1} + \frac{\nabla_\delta \mathcal{L}_{\text{DORMANT}}}{\text{mean}(\|\nabla_\delta \mathcal{L}_{\text{DORMANT}}\|)}$
  $\delta_i \leftarrow \delta_{i-1} + \gamma \cdot \text{sign}(g_i)$
  $\delta_i \leftarrow clip(\delta_i, -\eta, \eta)$
end for
$x_p \leftarrow x + \delta_N$
return x_p

The complete optimization objective of DORMANT is:

$$\begin{aligned} \underset{\|\delta\|_\infty \le \eta}{arg\,max} \mathcal{L}_{\text{DORMANT}} = \underset{\|\delta\|_\infty \le \eta}{arg\,max} & \lambda_1 \cdot \mathcal{L}_{vae} + \lambda_2 \cdot \mathcal{L}_{feature} \\ & + \lambda_3 \cdot \mathcal{L}_{frame} - \lambda_4 \cdot \mathcal{L}_{lpips}, \end{aligned} \tag{13}$$

where λs control the weight of each loss term. By default, we set these values as $\lambda_1 = 10$, $\lambda_2 = 100$[3], $\lambda_3 = 1$ and $\lambda_4 = 10$. The detailed optimization process is presented in Alg. 1.

4 Evaluation

4.1 Experimental Setup

Pose-driven Human Image Animation Methods. We comprehensively evaluate the effectiveness and transferability of DORMANT on 8 cutting-edge and widely-used human image animation methods, including Animate Anyone [30], MagicAnimate [74], MagicPose [13], MusePose [66], Champ [86], MuseV [73], UniAnimate [71], and ControlNeXt [53]. These 8 methods represent SOTA open-source animation techniques within the last two years and cover all major types. Despite advances in controllability and realism, some methods have also been implicated in generating dancing deepfakes on video platforms, raising societal concerns [15, 42, 52].

Datasets. We conduct experiments on 4 datasets: TikTok [33], Champ [86], UBC Fashion [78], and TED Talks [65]. For the TikTok dataset, we follow the settings used in prior human image animation works [13, 70] and utilize their 10 TikTok-style

[3] For $\mathcal{L}_{feature}$, we set $\lambda_{clip} = 10$ and $\lambda_{ref} = \lambda_2 = 100$, as ReferenceNet has demonstrated the ability to extract more dense and detailed features necessary for appearance alignment compared to the CLIP image encoder [30, 74].

Figure 3: Qualitative comparisons with baseline protections against various pose-driven human image animation methods.

videos showing different people from the web for evaluation. These videos contain between 248 and 690 frames. Specifically, we use the first frame of each video as the reference image and extract pose sequences from the remaining frames using DWPose [76] or DensePose [22] to guide the video generation. Additionally, we also sample 10 videos from the Champ training set, and the UBC Fashion and TED Talks test sets for evaluation, resulting in videos with 237-848, 303-355, and 135-259 frames, respectively. All video frames are resized to 512×512, and we further conduct experiments on other image resolutions, random reference images, and jump-cut pose sequences in Appendix A.

Baseline Protections. We compare the protection performance of DORMANT with 6 baseline methods defending against text/image-to-image generation (SDS [75], Glaze 2.1 [63], Mistv2 [84], PhotoGuard [62], and AntiDB [68]) and image-to-video generation (VGMShield [50]). For a fair comparison, we set the perturbation budget $\eta = 16/255$ and PGD iterations $N = 200$ for all protection methods[4]. We evaluate image similarity before and after protections in Appendix E.

Transformations and Purifications. We evaluate the robustness of DORMANT against 5 popular image transformations: JPEG compression, Gaussian blur, Gaussian noise, Median blur, and Bit squeeze. Additionally, we further conduct experiments on 6 advanced purification methods specifically designed to purify the added protective perturbation, including Impress [10], DiffPure [47], DDSPure [11], GrIDPure [83], Diffshortcut [44], and Noisy Upscaling [27].

Metrics. We evaluate the quality of generated videos using 6 image- and video-wise generative metrics used in prior human image animation methods [13, 70]. For the assessment of image-level quality, we report frame-wise LPIPS [79], Fréchet Inception Distance (FID) [24], Peak Signal to Noise Ratio (PSNR) [28], and Structural Similarity Index Measure (SSIM) [72]; while for video-level evaluation, we concatenate every consecutive 16 frames to form a sample and report FID-VID [6] and Fréchet Video Distance (FVD) [67], respectively.

Platform. All our experiments are conducted on a server running a 64-bit Ubuntu 22.04.4 system with Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz, 512GB memory, and four Nvidia A800 PCIe GPUs with 80GB memory.

4.2 Protection Performance

Overall Results. We evaluate the protection effectiveness and transferability of DORMANT against eight pose-driven human image animation methods on the TikTok dataset, with the quantitative results shown in Table 1. Evaluation across six metrics demonstrates significant degradation in the quality of generated videos due to the protective effect of DORMANT on all eight animation methods. Specifically, the average FID-VID, FVD, LPIPS, and FID increase from 48.03, 382.04, 0.30,

[4]Glaze is closed-source software whose perturbation budget cannot be exactly set to 16/255, we set the intensity to `High`, the highest setting.

Table 1: Quantitative comparisons with baseline protections against various pose-driven human image animation methods. ↑ indicates that a higher value of the metric signifies poorer video quality and thus better protection, while ↓ indicates the opposite.

Method	Metric	Clean	SDS [75]	Glaze [63]	Mistv2 [84]	VGMShield [50]	PhotoGuard [62]	AntiDB [68]	DORMANT
Animate Anyone [30]	FID-VID↑	41.61	108.40	109.98	59.49	53.95	112.98	84.75	**162.87**
	FVD↑	362.75	859.75	1301.96	678.56	625.83	994.84	1055.58	**1364.00**
	LPIPS↑	0.282	0.556	0.582	0.554	0.584	0.565	0.542	**0.639**
	FID↑	65.00	140.69	172.04	142.04	142.07	143.41	181.52	**245.06**
	PSNR↓	17.76	16.30	15.23	17.38	16.75	15.94	15.76	**11.82**
	SSIM↓	0.741	0.684	0.446	0.553	0.510	0.682	0.441	**0.374**
MagicAnimate [74]	FID-VID↑	37.04	148.59	92.20	66.49	56.00	145.01	60.61	**174.93**
	FVD↑	374.99	1051.55	1002.92	616.33	409.64	1081.05	620.48	**1174.52**
	LPIPS↑	0.268	0.409	0.531	0.499	0.506	0.421	0.502	**0.604**
	FID↑	68.86	131.53	132.25	119.88	111.03	127.21	115.07	**235.23**
	PSNR↓	18.29	15.03	15.74	17.33	17.72	14.78	17.18	**13.89**
	SSIM↓	0.758	0.676	0.493	0.595	0.600	0.675	0.535	**0.369**
MagicPose [13]	FID-VID↑	67.73	98.76	120.89	107.22	97.04	**134.99**	98.76	133.06
	FVD↑	431.39	902.56	1058.04	899.49	764.42	1013.40	902.56	**1413.94**
	LPIPS↑	0.291	0.523	0.525	0.505	0.517	0.434	0.523	**0.575**
	FID↑	55.06	132.95	130.18	119.35	105.56	116.43	132.95	**176.29**
	PSNR↓	17.24	16.06	15.67	16.21	16.46	15.58	16.06	**13.77**
	SSIM↓	0.745	0.506	0.510	0.534	0.550	0.681	0.506	**0.432**
MusePose [66]	FID-VID↑	30.17	73.77	107.58	52.87	47.93	92.29	73.77	**182.47**
	FVD↑	343.91	840.79	1208.92	556.01	441.20	800.16	840.79	**1265.38**
	LPIPS↑	0.284	0.524	0.574	0.532	0.560	0.462	0.524	**0.642**
	FID↑	66.08	147.22	163.89	126.60	128.12	126.95	147.22	**280.57**
	PSNR↓	17.77	16.08	15.12	17.46	17.42	16.67	16.08	**11.50**
	SSIM↓	0.745	0.481	0.461	0.601	0.572	0.712	0.481	**0.381**
Champ [86]	FID-VID↑	85.19	144.27	148.12	107.79	97.79	152.46	111.93	**158.70**
	FVD↑	498.01	1017.24	**1291.57**	806.41	779.76	1119.71	1062.11	1234.56
	LPIPS↑	0.394	0.498	0.632	0.605	0.625	0.510	0.600	**0.655**
	FID↑	71.66	126.57	153.60	123.60	130.77	130.89	167.54	**233.61**
	PSNR↓	13.29	13.68	11.79	13.03	12.62	13.63	12.00	**11.57**
	SSIM↓	0.642	0.625	0.391	0.496	0.470	0.621	0.415	**0.349**
MuseV [73]	FID-VID↑	60.75	128.61	98.63	74.06	89.16	119.82	89.46	**151.46**
	FVD↑	430.96	820.45	987.90	601.51	603.88	852.04	622.80	**1078.03**
	LPIPS↑	0.310	0.563	0.543	0.533	0.539	0.512	0.407	**0.607**
	FID↑	82.19	127.12	139.48	115.29	117.65	115.89	136.73	**240.82**
	PSNR↓	16.11	15.79	15.22	16.32	15.78	15.98	15.36	**14.48**
	SSIM↓	0.721	0.665	0.563	0.649	0.644	0.681	0.675	**0.526**
UniAnimate [71]	FID-VID↑	26.03	86.59	105.43	49.75	43.15	81.76	56.77	**113.73**
	FVD↑	277.02	704.07	**1134.13**	572.21	490.22	667.84	708.05	1031.22
	LPIPS↑	0.253	0.521	0.579	0.566	0.567	0.488	0.505	**0.580**
	FID↑	52.63	131.51	153.04	115.08	119.63	123.60	167.66	**250.30**
	PSNR↓	18.76	15.76	15.58	17.91	17.98	16.09	17.40	**15.57**
	SSIM↓	0.762	0.680	0.474	0.566	0.587	0.709	0.545	**0.470**
ControlNeXt [53]	FID-VID↑	35.73	105.07	96.45	72.94	84.25	80.65	79.35	**110.37**
	FVD↑	337.31	1001.21	1134.70	764.49	862.51	775.46	839.82	**1141.36**
	LPIPS↑	0.301	0.478	0.501	0.494	**0.541**	0.428	0.492	0.538
	FID↑	62.87	130.81	126.56	106.19	140.87	112.98	138.89	**163.06**
	PSNR↓	16.88	**11.95**	13.88	14.44	13.92	13.55	14.85	13.19
	SSIM↓	0.734	0.529	0.505	0.547	0.485	0.617	0.532	**0.441**

and 65.54 to 148.45, 1212.88, 0.61 and 228.12, and the average PSNR and SSIM decrease from 17.01 and 0.731 to 13.22 and 0.418, respectively. We show non-cherry-picked generated video frames in Figure 3, which display mismatched identities, distorted backgrounds, and visual artifacts. These results highlight the effectiveness of DORMANT in preventing unauthorized image usage in human video generation, thereby safeguarding individuals' rights to portrait and privacy.

We also compare DORMANT's protection performance with six baseline methods that defend against unauthorized image and video generation. Quantitative results in Table 1 and qualitative results in Figure 3 demonstrate the superior protection performance of DORMANT, showing the greatest effectiveness in degrading the quality of generated videos. While Dormant exhibits the strongest transferability and performs effectively across all 8 animation techniques, baseline protections show limited transferability and are only effective for certain animation methods, yielding some better metrics with different protective effects. Specifically, the concurrent work VGMShield achieves slightly higher LPIPS on the SVD-based method ControlNeXt but lacks transferability to other animation methods. SDS tends to lighten colors and smooth out details, resulting in better PSNR on ControlNeXt. PhotoGuard induces noticeable inconsistencies between adjacent frames, showing slightly higher FID-VID on MagicPose. Glaze produces visual artifacts and intensifies colors in generated videos, leading to better FVD on Champ and UniAnimate. Moreover, although these baseline protections may perform better on some metrics for certain animation methods, DORMANT still 1) shows a clear advantage in other metrics for the corresponding animation method; and 2) outperforms them across all other animation methods.

Figure 4: Qualitative results on various datasets.

Results on Various Datasets. Besides the TikTok dataset, we also evaluate the protection performance of DORMANT on three other datasets using Animate Anyone, including human dance generation on the Champ dataset, fashion video synthesis on the UBC Fashion dataset, and speech video generation on the TED Talks dataset. Quantitative results in Table 2 demonstrate the remarkable performance of DORMANT in defending against pose-driven human image animation, with the average FID-VID, FVD, LPIPS and FID increasing from 53.78, 406.03, 0.22 and 60.81 to 166.28, 1288.20, 0.57 and 260.53, and the average PSNR and SSIM decreasing from 19.44 and 0.745 to 14.69 and 0.486, respectively. The poor-quality video frames shown in Figure 4 also validate the protection effectiveness of DORMANT.

4.3 Human and GPT-4o Studies

Human Study. To bridge the gap between quantitative metrics and human perception, as well as study whether DORMANT outperforms baseline methods in protecting privacy and portrait rights from the perspective of real-world audiences, we conduct a survey study. Our questionnaire consists of two parts: 1) demographic data collection, including five questions on age, gender, education, expertise, and familiarity; and 2) ten questions asking participants to select the video that best protects portrait and privacy rights compared to the reference, each question containing a reference video and five videos generated from images protected by DORMANT and four baseline methods (SDS, Glaze, VGMShield, and PhotoGuard). The videos used are the 10 videos in the TikTok test set, derived from experiments conducted using Animate Anyone, with the corresponding metrics provided in Table 1. The detailed questionnaire can be found in Appendix F.1.

We distributed the questionnaire online and recruited participants via social media, offering lottery-based compensation upon completion. A total of 82 valid responses were collected. The participants' ages primarily ranged from 18 to 54, with the majority (63.41%) falling between 25 and 34. Among the participants, 57.32% were male and 40.24% were female; 80.48% held or were pursuing a bachelor's degree or higher; 39.02% had a background in computer-related fields; and 70.73% had at least heard of image-to-video generation or were more familiar with it. As shown in Table 3, DORMANT achieved an overall pick rate of 74.27% (609/820), significantly outperforming baseline protections. Detailed results for each question are shown in Figure 15(a) in Appendix F.3. Notably, 67 participants selected DORMANT for at least five questions, including 29 with a computer-related background (32 in total). 37 participants selected DORMANT for all ten questions, while only 8 participants did not choose DORMANT for any of the ten questions, with SDS being their preferred method, yielding a pick rate of 72.50% (58/80). Of these eight participants, only one had heard of image-to-video generation, and the others were less familiar with it.

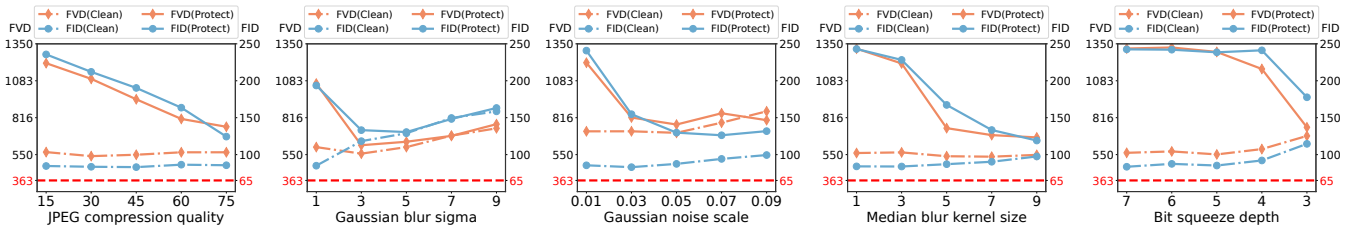
Table 2: Protection performance on various datasets.

Dataset		FID-VID↑	FVD↑	PSNR↓	SSIM↓	LPIPS↑	FID↑
TikTok [33]	**Clean**	41.61	362.75	17.76	0.741	0.282	65.00
	Protect	162.87	1364.00	11.82	0.374	0.639	245.06
Champ [86]	**Clean**	45.92	379.64	18.04	0.661	0.286	56.72
	Protect	156.03	1465.48	13.72	0.397	0.597	252.03
UBC Fashion [78]	**Clean**	38.04	329.28	20.39	0.876	0.086	33.74
	Protect	78.67	780.96	18.38	0.692	0.432	164.99
TED Talks [65]	**Clean**	89.54	552.45	21.56	0.701	0.233	87.78
	Protect	267.54	1542.37	14.82	0.479	0.627	380.04

Table 3: Results of human and GPT-4o studies.

Metric	SDS	Glaze	VGMShield	PhotoGuard	DORMANT
Human Pick Rate↑	10.12%	9.51%	4.27%	1.83%	**74.27%**
GPT-4o Rank↓	4.0	2.0	2.7	4.8	**1.5**

GPT-4o Study. As LLM-as-a-Judge is now widely adopted in many complex tasks [21], we also use GPT-4o [48] to simulate the above human study. GPT-4o is a multimodal large language model capable of understanding image semantics and perceiving pixel-level differences. We input GPT-4o with a reference and five generated frame sequences, asking it to rank them based on their effectiveness in protecting portrait and privacy rights, focusing mainly on factors such as appearance mismatches and background changes that differentiate the generated frames from the reference. GPT-4o is instructed to analyze step by step and output both the ranking and its reasoning. Our prompt is detailed in Appendix F.2. We further manually verify all outputs from GPT-4o to ensure there are no hallucinations. As shown in Table 3, DORMANT achieves the highest average ranking of 1.5, and is consistently ranked first or second across all samples (see Figure 15(b) in Appendix F.3). While GPT-4o ranks DORMANT first, it notes that this method "makes the individual unrecognizable and drastically alters the background". Glaze is preferred by GPT-4o when DORMANT is ranked second, as it "incorporates more prominent and deliberate texture patterns".



(a) JPEG compression (b) Gaussian blur (c) Gaussian noise (d) Median blur (e) Bit squeeze

Figure 5: Protection robustness of DORMANT against various transformations under different parameter settings.

Figure 6: Qualitative results of transformations that most degrade the protection, with JPEG compression (quality=75), Gaussian blur (sigma=3), Gaussian noise (scale=0.05), Median blur (kernel size=9) and Bit squeeze (depth=3).

4.4 Protection Robustness

Robustness against Transformations. We evaluate five transformations commonly used in prior studies [17,27,83], including three transformations applied in EoT (JPEG compression, Gaussian blur, and Gaussian noise) and two unseen transformations (Median blur and Bit squeeze). We apply them with different parameter settings to both the original and protected images to investigate their impact on the quality of generated videos and the effectiveness of protection. As shown in Figure 5 (with the red dotted line representing the results of videos generated using the original image without transformations), the results indicate that: 1) transformations can only reduce the protection effect to some extent but cannot completely remove the impact of the protective perturbation; 2) transformations also affect the quality of generated videos. In particular, increasing the transformation parameters to minimize the protection effect would also noticeably degrade the quality of the generated videos. Figure 6 presents the qualitative results when various transformations most degrade the protection effect. Even in these cases, the generated videos still exhibit low quality due to residual protective perturbation and the impact of excessive transformations, displaying issues including distorted backgrounds, mismatched faces, and visual artifacts. These results validate the robustness of DORMANT against various transformations.

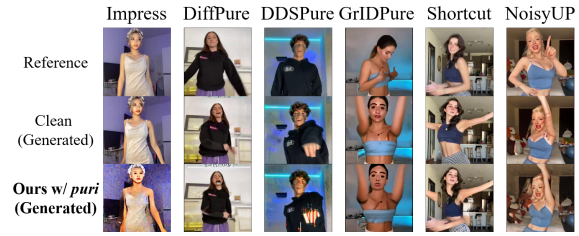


Figure 7: Qualitative results against various purifications.

Table 4: Quantitative results against various purifications.

Method	FID-VID↑	FVD↑	PSNR↓	SSIM↓	LPIPS↑	FID↑
Impress [10]	171.49	1386.19	13.31	0.341	0.628	259.26
DiffPure [47]	54.43	490.63	17.57	0.712	0.337	94.23
DDSPure [11]	71.38	762.01	17.14	0.687	0.398	126.93
GrIDPure [83]	62.69	465.87	16.78	0.670	0.338	101.40
Diffshortcut [44]	73.48	605.54	16.11	0.545	0.514	152.60
Noisy Upscaling [27]	78.10	673.03	17.02	0.680	0.414	144.24
Clean	**41.61**	**362.75**	**17.76**	**0.741**	**0.282**	**65.00**

Robustness against Purifications. We evaluate the robustness of DORMANT against six purification methods, and the results are shown in Table 4. While the protective effect of DORMANT is somewhat reduced by these purifications, there still remains a noticeable quality gap between the resulting videos and those generated from the original images. Qualitative results are presented in Figure 7. Impress shows almost no purifying effect. DiffPure and DDSPure would introduce external strange patterns and light streaks into the generated videos. GrIDPure and Noisy Upscaling fail to restore detailed facial features. In contrast, while Diffshortcut utilizes CodeFormer [85] to restore human faces (which are vivid but differ from the reference), there are still noticeable artifacts in other regions. As a result, videos generated using protected images after purification cannot be used directly for malicious purposes due to their low quality, and the individuals' privacy and portrait rights are still effectively safeguarded.

4.5 Protection Transferability

Results on Image-to-Video/Image. Besides human image animation methods that use pose sequences to guide video

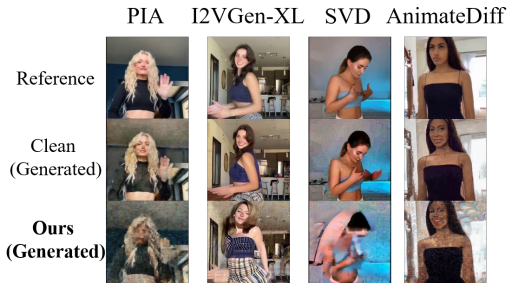


Figure 8: Qualitative results on image-to-video methods.

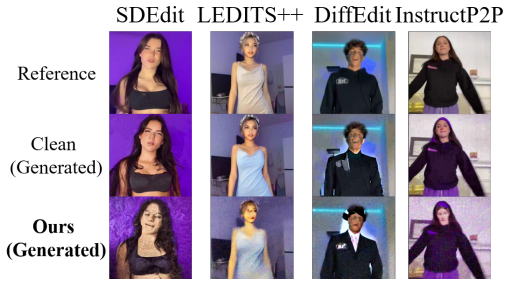


Figure 9: Qualitative results on image-to-image methods.

generation from the reference image, we conduct additional experiments on four other types of image-to-video methods that directly generate videos from the given image without external guidance (SVD) and those that use text prompts for guidance (PIA, I2VGen-XL, and AnimateDiff). Moreover, we also consider the scenario of image manipulation and evaluate the performance of DORMANT using four image-to-image methods that use text prompts as conditions for image editing. We utilize BLIP-2 [37] to generate captions corresponding to the reference images, resulting in prompts such as "A woman in a white dress". To guide video generation, we add the suffix "*is dancing*" to these captions. For image-to-image methods, we use the original captions for SDEdit and InstructPix2Pix, while for LEDITS++ and DiffEdit, we generate variant prompts using GPT-4o, such as "A woman in a *blue* dress". Since FID-VID and FVD cannot be calculated here, we employ two other metrics, CLIP-I and DINO, which compute the average pairwise cosine similarity between CLIP image embeddings [54] and ViT-S/16 DINO embeddings [12] of generated and reference images, respectively.

As shown in Table 5, DORMANT demonstrates excellent protection performance, resulting in significant degradation in the quality of generated videos and images. For the four image-to-video methods, the average CLIP-I, DINO, and SSIM decrease from 0.771, 0.784, and 0.559 to 0.638, 0.418, and 0.342, and the average LPIPS and FID increase from 0.501 and 136.57 to 0.677 and 294.09, respectively. While for the four image-to-image methods, the average CLIP-I, DINO, and SSIM decrease from 0.779, 0.827, and 0.684 to 0.655, 0.521, and 0.410, and the average LPIPS and FID increase

Figure 10: Qualitative results on commercial services.

Table 5: Quantitative results on various image-to-video and image-to-image methods.

Task	Method		CLIP-I↓	DINO↓	SSIM↓	LPIPS↑	FID↑
Image-to-Video	PIA [81]	Clean	0.752	0.768	0.573	0.519	142.69
		Protect	0.655	0.464	0.409	0.693	275.74
	I2VGen-XL [80]	Clean	0.779	0.730	0.467	0.560	158.72
		Protect	0.641	0.414	0.284	0.690	300.19
	SVD [7]	Clean	0.799	0.805	0.540	0.476	124.69
		Protect	0.658	0.377	0.317	0.683	321.32
	AnimateDiff [23]	Clean	0.752	0.831	0.655	0.447	120.18
		Protect	0.596	0.416	0.359	0.642	279.09
Image-to-Image	SDEdit [46]	Clean	0.733	0.841	0.609	0.442	140.98
		Protect	0.599	0.535	0.305	0.634	264.35
	LEDITS++ [8]	Clean	0.823	0.853	0.735	0.288	119.50
		Protect	0.682	0.591	0.488	0.626	222.32
	DiffEdit [16]	Clean	0.788	0.843	0.738	0.267	180.22
		Protect	0.665	0.430	0.428	0.649	375.78
	InstructPix2Pix [9]	Clean	0.773	0.772	0.653	0.391	156.13
		Protect	0.675	0.529	0.420	0.686	275.54

from 0.347 and 149.21 to 0.649 and 284.50, respectively. Visualization results are shown in Figure 8 and Figure 9, where both the generated video frames and images exhibit poor quality. These results further highlight the effectiveness and transferability of DORMANT. The features contained in the reference image are effectively and comprehensively disrupted by the protective perturbation from the perspective of various LDMs used in these generation methods, which utilize different architectures and are trained on different datasets.

Results on Commercial Services. We further evaluate the performance of DORMANT on six real-world commercial services. To effectively defend against these closed-source generative models in a black-box manner, we set the budget η to $32/255$ to generate protective perturbation. The results are presented in Table 6. DORMANT significantly degrades the quality of generated videos and images, with the average CLIP-I, DINO, and SSIM decreasing from 0.804, 0.788, and 0.630 to 0.662, 0.521, and 0.475, and the average LPIPS and FID increasing from 0.474 and 145.22 to 0.643 and 269.64, respectively. As shown in Figure 10, generated videos and images suffer from poor quality, such as mismatched human appearance (Gen-2, DreaMoving, and Ying) and distorted visuals (Viva, NovelAI, and Scenario). These results validate that DORMANT effectively safeguards portrait and privacy rights, showcasing its capability for real-world applications in preventing human image misuse.

Table 6: Quantitative results on various commercial services.

Task	Commercial Service		CLIP-I↓	DINO↓	SSIM↓	LPIPS↑	FID↑
Image-to-Video	Viva [25]	Clean	0.880	0.888	0.717	0.359	82.94
		Protect	0.652	0.486	0.418	0.642	308.24
	Gen-2 [60]	Clean	0.778	0.752	0.633	0.477	150.84
		Protect	0.661	0.604	0.543	0.655	236.08
	DreaMoving [20]	Clean	0.805	0.764	0.539	0.551	138.50
		Protect	0.664	0.563	0.470	0.661	215.73
	Ying [2]	Clean	0.847	0.729	0.535	0.605	170.39
		Protect	0.677	0.472	0.407	0.681	278.24
Image-to-Image	NovelAI [4]	Clean	0.710	0.724	0.654	0.493	186.74
		Protect	0.652	0.532	0.518	0.591	243.71
	Scenario [32]	Clean	0.803	0.870	0.702	0.360	141.91
		Protect	0.663	0.466	0.493	0.626	335.85

4.6 Ablation Study

Impact of Proposed Objectives. To study the impact of $\mathcal{L}_{vae}$ and $\mathcal{L}_{feature}$ (feature misextraction), $\mathcal{L}_{frame}$ (frame incoherence), we conduct experiments by removing each corresponding loss while keeping the rest of $\mathcal{L}_{\text{DORMANT}}$ unchanged. As shown in Figure 11, the absence of any of these three losses results in a degradation of the protection effect, highlighting their essential roles in the overall effectiveness of the protection. Notably, $\mathcal{L}_{feature}$ has the most significant impact, which is designed to induce misextraction of appearance features from the reference image using CLIP and ReferenceNet. This result aligns with our analysis in Section 2.3, where we emphasize that the main usage of the reference image in pose-driven human image animation is to extract its appearance features to guide video generation. Consequently, as shown in Table 1, existing protections, which lack specific designs to effectively disrupt these features, exhibit limited efficacy.

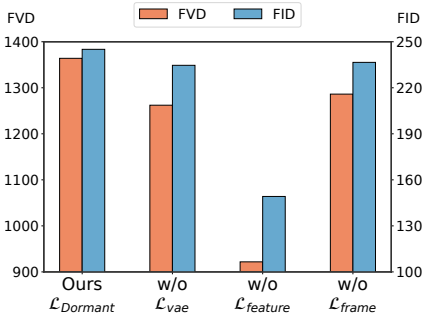
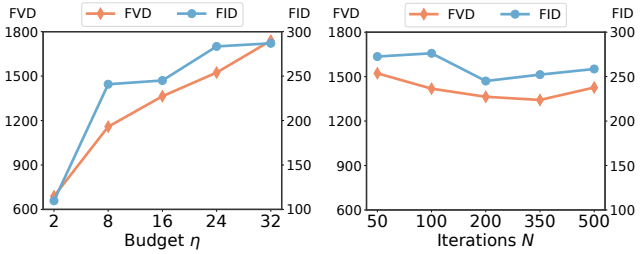


Figure 11: Ablation study on the proposed objectives.

We further study the impact of using other feature extractors in $\mathcal{L}_{feature}$. As mentioned in Section 3.2, we use the CLIP image encoder to extract semantic features and ReferenceNets to capture fine-grained details from the image. We replace our design of CLIP + ReferenceNets with DINO [12] + BLIP image encoder [38], yielding FVD and FID values of 1080.09 and 153.55 for the generated videos. This result outperforms most baseline protections (except Glaze) but falls short of



(a) Perturbation budget η (b) PGD iterations N

Figure 12: Impact of budget η and iterations N.

DORMANT by a noticeable margin, further highlighting the importance of inducing misextraction of appearance features and also the superiority of our design for feature extractors.

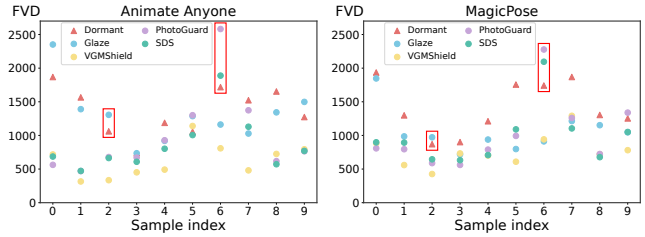
Impact of Perturbation Budget η and PGD Iterations N. We conduct experiments by varying η from $2/255$ to $32/255$ and N from 50 to 500, to investigate their impact on protection performance. The results are presented in Figure 12. As the perturbation budget η increases, the protection effect of DORMANT becomes progressively stronger, and $\eta = 8/255$ can already significantly degrade the quality of generated videos, resulting in FVD and FID values of 1158.46 and 240.83, respectively. This provides users with the flexibility to balance invisibility and effectiveness according to their needs. The increase in PGD iterations N initially causes the protection effect to decrease and then increase, possibly due to overfitting, which leads the optimization to fall into poor local maxima and then escape. Optimizing for 50 iterations is sufficient to significantly degrade the quality of generated videos, with FVD and FID being 1522.35 and 272.61, thereby further reducing the time cost required for protection. Impact of pose repeats F and decay factor μ is provided in Appendix C.

5 Discussion

Adaptive Attack. We propose an adaptive attack under the assumption that the attacker has access to multiple protected images of the victim. While animation methods typically take a single reference image as input to generate videos, the attacker can utilize these images to craft a refined reference. Inspired by [51] which reveals Glaze's vulnerability to linear interpolation and pixel averaging, we design a strategy to create a purified image from multiple protected images. Specifically, we apply linear interpolation to adjacent pairs of five protected images spaced five frames apart, followed by pixel averaging of the four interpolated results to produce the final image for video generation, with the resulting FID-FVD, FVD, LPIPS, FID, PSNR, and SSIM on the TikTok dataset being 69.18, 728.89, 0.474, 162.50, 17.06, and 0.641, respectively. While the protective effect is partially reduced, the generated videos remain of noticeably low quality. More adaptive attacks are provided in Appendix B.

Application Scenarios. DORMANT applies protective perturbations to human images to prevent potential unauthorized usage in pose-driven animation methods. It is primarily targeted at users who: 1) may place human images in uncontrollable environments like social media or the web; and 2) prioritize privacy and portrait rights and accept sacrificing a bit of image quality for protection. DORMANT may not be directly suitable for legitimate use cases that involve processing or editing original images for valid purposes, such as photo retouching or authorized content creation.

Anomalous Cases. We examine cases where baseline protections outperform DORMANT across at least four animations on the TikTok dataset. While DORMANT exhibits sufficient effectiveness, these methods yield even better protective effects on certain samples. Specifically, Glaze performs better on sample 002, producing highly exaggerated colors and visible speckles in the generated videos. PhotoGuard and SDS perform better on sample 006 which features sharply contrasting colors, by lightening the colors and blurring sharpness and details in the generated videos. We present the FVD for experiments on Animate Anyone and MagicPose as examples in Figure 13, and qualitative results in Figure 15 in Appendix D.



(a) Results of Animate Anyone (b) Results of MagicPose

Figure 13: Results of experiments on Animate Anyone and MagicPose, with red boxes marking samples where baseline protections perform better across both animation methods.

We also examine failure cases where DORMANT performs relatively poorly against the commercial service Gen-2 and the latest Gen-3. As shown in Figure 16 in Appendix D, DORMANT mainly induces identity mismatch but introduces minimal distortion, and the generated videos remain realistic. Image-to-video models applied in commercial services are typically trained on larger, higher-quality datasets with more parameters. Exploring how to get feedback from these advanced models via black-box access to improve DORMANT's transferability would be meaningful for future work.

Limitations and Future Works. A key limitation of DORMANT is that its effectiveness is validated based on empirical evaluation, rather than being provably secure. Additionally, DORMANT faces the challenge of being future-proof, as image-to-video generation is continuously evolving. While DORMANT has shown effectiveness and is expected to remain effective against existing feature extractors used in current open-source SOTA animation methods that are driven by pose, its performance against future paradigms in feature extraction and different video generation techniques guided by other modal conditions requires ongoing validation and improvements. Future work could explore the use of expert models to specifically extract disentangled image features (*e.g.*, human face, body posture, and background) to induce feature misextraction, and integrate guidance from additional modalities to create frame incoherence during optimization.

6 Related Work

LDM for Unauthorized Image Usage. Latent diffusion models [57] have demonstrated a remarkable capability to create authentic-looking images, which also raises significant concerns about unauthorized image usage. Malicious attackers can easily mimic an artist's style [29, 59] or edit images into new contexts [8, 9, 16, 46] without consent. Beyond image generation, unauthorized images can also be used to create fabricated videos [7, 23, 80, 81]. In particular, pose-driven human image animation [13, 30, 36, 53, 66, 69–71, 73, 74, 82, 86] can generate malicious videos by animating human images using pose sequences as guidance.

Protection Methods and Countermeasures. Previous protection methods apply protective perturbations to images to prevent unauthorized usage in DNN-based recognition [14, 39, 64] and GAN-based manipulation [31, 40, 77]. However, these methods have shown limited effectiveness against advanced LDMs due to their distinct backbone networks [41, 63]. To defend against LDM-based text/image-to-image generation, recent research has proposed protections targeting customization techniques [68, 84], VAE [62, 63], and denoising UNet [41, 75]. The only concurrent work addressing unauthorized LDM-based image-to-video generation is VGMShield [50], which targets the image and video encoders of SVD. There also exists countermeasures [3, 10, 27, 44, 83] that scrutinize the effectiveness and robustness of current protection methods, proposing various techniques to circumvent them by purifying the added protective perturbation.

7 Conclusion

In this paper, we propose DORMANT, a novel approach designed to prevent unauthorized image usage in pose-driven human image animation. By applying the protective perturbation to the human image, DORMANT effectively degrades the quality of generated videos, resulting in mismatched appearances, visual distortions, and frame inconsistencies. To optimize the protective perturbation, we develop specific objective functions aimed at inducing misextraction of appearance features and causing incoherence among generated video frames. Experimental results demonstrate the effectiveness, transferability, and robustness of DORMANT, serving as a powerful tool for safeguarding portrait and privacy rights.

Acknowledgments

We thank our shepherd and all the anonymous reviewers for their constructive feedback. The IIE authors are supported in part by NSFC (92270204, U24A20236), CAS Project for Young Scientists in Basic Research (Grant No. YSBR-118).

Ethics Considerations

DORMANT is a defense approach designed to prevent image misuse in pose-driven human image animation methods, thereby safeguarding individuals' rights to portrait and privacy. To ensure fair and reproducible evaluations of the performance of DORMANT, we conducted experiments on commonly used public datasets, following the setups outlined in prior animation methods [13, 70, 74, 86]. All experimental data, including the fake videos generated during the experiments, were stored on private, secure servers and are strictly intended for academic research purposes only; they will not be shared beyond their intended scope.

During the research process, three experiments posed potential risks of unauthorized exposure of generated fake videos: the human study, the GPT-4o study, and experiments on commercial services. These experiments used the TikTok test set [33, 70], which permits research purposes. Although this publicly available dataset is widely used in many studies, obtaining explicit consent from the individuals depicted in the photos remains challenging. An alternative approach could have involved collecting and using data from volunteers who could explicitly consent. However, the terms of service for video generation platforms typically allow the use of user inputs and outputs for training and improving their AI models. While publicly available data may have already been integrated into these platforms' model training, uploading newly collected images from consenting individuals would introduce their data for the first time, potentially exposing them to additional privacy risks due to uncontrolled usage. To mitigate these risks and ensure fairness and reproducibility in our evaluation, we chose to use publicly available data.

Meanwhile, we remained vigilant about the potential exposure of generated fake videos and implemented precautionary measures to minimize risks. All participants in the human study were required to consent to not disseminate any content from the questionnaire, which was accessible for a limited time. Before uploading data to GPT-4o and video generation websites, we configured our account privacy settings to the strictest levels, including disabling "Chat History & Training" in GPT-4o, setting repositories to private, and upgrading to memberships with enhanced privacy features, *etc.* During the experiments, we documented the setups for reproducibility, downloaded and stored the generated results on private servers, and cleared all data from the accounts upon completion. All activities were conducted in full compliance with the usage rules and regulations of the respective platforms.

We conducted a human study to evaluate the performance of DORMANT from the perspective of human perception. The study was reviewed and received IRB approval at our institute. Before providing consent, participants were fully informed of the study's purpose, procedures, potential risks and benefits, and lottery-based compensation. They were also warned that the questionnaire might include unfiltered videos that could be disturbing, and were shown sample videos upon request if they expressed concern. Participants were allowed to withdraw from the study at any point, with any data they provided excluded from the analysis. Additionally, participants were required to consent to not disseminate any content from the questionnaire. Explicit consent was obtained through a digital agreement before participants gained access to the questionnaire, which was available for a limited duration. All participants remained anonymous. Demographic data, including gender, age, education, expertise, and familiarity, were collected solely for research purposes through single-choice questions with generalized options, including a "prefer not to say" option. No personally identifiable information was collected, and all data were securely and privately stored.

Open Science

We are committed to the principles of open science and have released our artifacts, including the source code of DORMANT and the evaluated test data at https://zenodo.org/records/14725876. The latest version of the code will be maintained and updated at https://github.com/Manu21JC/Dormant. While we primarily present visualizations of experimental results via figures in this paper, we do not plan to provide public links to these generated fake videos to mitigate potential risks of exposure and misuse, as discussed in Ethics Considerations. These videos are available upon request for research purposes only, with the condition that requestors agree not to share the videos beyond their intended scope.

References

[1] Stability AI. sd-vae-ft-mse, 2022. https://huggingface.co/stabilityai/sd-vae-ft-mse.

[2] Zhipu AI. Ying, 2024. https://chatglm.cn/video.

[3] Shengwei An, Lu Yan, Siyuan Cheng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, and Xiangyu Zhang. Rethinking the invisible protection against unauthorized image usage in stable diffusion. In *33nd USENIX Security Symposium (USENIX Security 24)*, 2024.

[4] Anlatan. Novelai, 2022. https://novelai.net/.

[5] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[6] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, 2019.

[7] 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.

[8] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024.

[9] 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.

[10] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36, 2024.

[11] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[13] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2024.

[14] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[15] Devin Coldewey. 'animate anyone' heralds the approach of full-motion deepfakes, 2023. https://techcrunch.com/2023/12/04/animate-anyone-heralds-the-approach-of-full-motion-deepfakes/.

[16] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023.

[17] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. Advertorch v0. 1: An adversarial robustness toolbox based on pytorch. *arXiv preprint arXiv:1902.07623*, 2019.

[18] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[19] 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.

[20] Alibaba Group. Dreamoving, 2023. https://www.modelscope.cn/studios/vigen/video_generation/summary.

[21] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[22] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] 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.

[25] HiDream.ai. Viva, 2023. https://vivago.ai/explore.

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

[27] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*, 2024.

[28] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[29] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[30] 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.

[31] Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Initiative defense against facial manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[32] Scenario Inc. Scenario, 2023. https://www.scenario.com/.

[33] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021.

[34] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014.

[35] Lambda Labs. Stable diffusion image variations, 2022. https://huggingface.co/lambdalabs/sd-image-variations-diffusers.

[36] Wentao Lei, Jinting Wang, Fengji Ma, Guanjie Huang, and Li Liu. A comprehensive survey on human video generation: Challenges, methods, and insights. *arXiv preprint arXiv:2407.08428*, 2024.

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[39] Yuezun Li, Xin Yang, Baoyuan Wu, and Siwei Lyu. Hiding faces in plain sight: Disrupting ai face synthesis with adversarial perturbations. *arXiv preprint arXiv:1906.09288*, 2019.

[40] Zheng Li, Ning Yu, Ahmed Salem, Michael Backes, Mario Fritz, and Yang Zhang. {UnGANable}: Defending against {GAN-based} face manipulation. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 7213–7230, 2023.

[41] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pages 20763–20786. PMLR, 2023.

[42] Fazhong Liu, Yan Meng, Tian Dong, Guoxing Chen, and Haojin Zhu. Detection and attribution of diffusion model of character animation based on spatio-temporal attention. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pages 105–108, 2024.

[43] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

[44] Yixin Liu, Ruoxi Chen, and Lichao Sun. Investigating and defending shortcut learning in personalized diffusion models. *arXiv preprint arXiv:2406.18944*, 2024.

[45] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[46] 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*, 2022.

[47] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827. PMLR, 2022.

[48] OpenAI. Gpt-4o, 2024. https://chatgpt.com/.

[49] OpenAI. Sora, 2024. https://sora.com/.

[50] Yan Pang, Yang Zhang, and Tianhao Wang. Vgmshield: Mitigating misuse of video generative models. *arXiv preprint arXiv:2402.13126*, 2024.

[51] Josephine Passananti, Stanley Wu, Shawn Shan, Haitao Zheng, and Ben Y Zhao. Disrupting style mimicry attacks on video imagery. *arXiv preprint arXiv:2405.06865*, 2024.

[52] Gan Pei, Jiangning Zhang, Menghan Hu, Guangtao Zhai, Chengjie Wang, Zhenyu Zhang, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024.

[53] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

[54] 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.

[55] Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, and Florian Tramer. Data poisoning won't save you from facial recognition. In *International Conference on Learning Representations*, 2022.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] Runway. Gen-2, 2023. https://runwayml.com/research/gen-2.

[61] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[62] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *International Conference on Machine Learning*, pages 29894–29918. PMLR, 2023.

[63] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

[64] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020.

[65] 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.

[66] Zhengyan Tong, Chao Li, Zhaokang Chen, Bin Wu, and Wenjiang Zhou. Musepose: a pose-driven image-to-video framework for virtual human generation, 2024. https://github.com/TMElyralab/MusePose.

[67] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[68] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023.

[69] Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. Vividpose: Advancing stable video diffusion for realistic human image animation. *arXiv preprint arXiv:2405.18156*, 2024.

[70] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024.

[71] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024.

[72] 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.

[73] Zhiqiang Xia, Zhaokang Chen, Bin Wu, Chao Li, Kwok-Wai Hung, Chao Zhan, Yingjie He, and Wenjiang Zhou. Musev: Infinite-length and high fidelity virtual human video generation with visual conditioned parallel denoising, 2024. https://github.com/TMElyralab/MuseV.

[74] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024.

[75] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2024.

[76] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023.

[77] Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 53–62, 2020.

[78] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff,UK, September 9-12, 2019*, page 51. BMVA Press, 2019.

[79] 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.

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

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

[82] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

[83] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24398–24407, 2024.

[84] Boyang Zheng, Chumeng Liang, Xiaoyu Wu, and Yan Liu. Improving adversarial attacks on latent diffusion model. *arXiv preprint arXiv:2310.04687*, 2023.

[85] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

[86] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024.

Appendix

A Protection Generality

Results on Various Resolutions of Images. As mentioned in Section 4.1, our experiments are mainly conducted on 512×512 images. Here we investigate the protection performance of DORMANT on four other resolutions: $\{256 \times 256, 256 \times 512, 512 \times 768, 768 \times 768\}$. As shown in Table 7, DORMANT exhibits excellent protection performance across all image resolutions, with the average FID-VID, FVD, LPIPS and FID increasing from 53.83, 447.50, 0.31 and 81.01 to 215.96, 1628.55, 0.64 and 251.09, and the average PSNR and SSIM decreasing from 16.85 and 0.701 to 11.85 and 0.370, respectively. These results highlight the applicability of DORMANT for protecting human images of different resolutions.

Table 7: Quantitative results on various resolutions of images.

Resolution		FID-VID↑	FVD↑	PSNR↓	SSIM↓	LPIPS↑	FID↑
256 × 256	Clean	56.23	462.36	16.04	0.625	0.304	94.07
	Protect	260.26	2165.94	13.19	0.306	0.596	285.03
256 × 512	Clean	69.54	480.56	15.74	0.637	0.349	83.85
	Protect	289.11	1769.24	12.06	0.322	0.640	292.09
512 × 512	Clean	41.61	362.75	17.76	0.741	0.282	65.00
	Protect	162.87	1364.00	11.82	0.374	0.639	245.06
512 × 768	Clean	58.86	505.83	16.84	0.722	0.328	85.67
	Protect	204.48	1419.64	11.62	0.403	0.653	243.46
768 × 768	Clean	42.93	425.98	17.89	0.779	0.284	76.45
	Protect	163.07	1423.93	10.58	0.441	0.650	189.82

Results on Random References and Jump-cut Poses. As mentioned in Section 4.1, we use the first frame of each video as the reference image and extract pose sequences from the remaining frames for video generation. We further conduct two additional experiments to examine whether altering these settings would impact DORMANT's effectiveness. First, we randomly select a non-first frame from each video to serve as the reference image, with the resulting FID-FVD, FVD, LPIPS, FID, PSNR, and SSIM on the TikTok dataset being 162.83, 1352.56, 0.633, 225.50, 11.97, and 0.401, respectively. Second, we simulate jump-cut pose sequences by combining the first half of the pose sequence from one video with the second half from another, yielding CLIP-I, DINO, SSIM, LPIPS, and FID values of 0.609, 0.437, 0.404, 0.682, and 247.73, respectively. These results show that DORMANT's protection performance remains robust under both variations in the reference image and pose sequence settings.

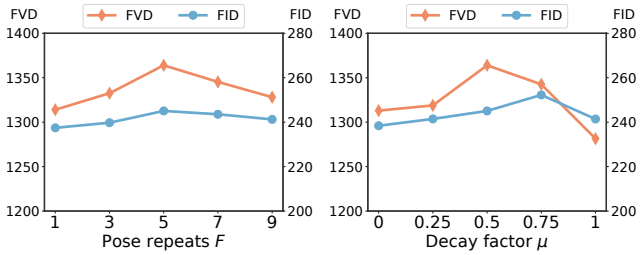
B More Adaptive Attacks

Robustness against Robust Fine-tuning. Inspired by [55] which reveals that robust training with protected images and their corresponding ground-truth labels could defeat protections against face recognition models [14, 64], we investigate whether this strategy would affect DORMANT's performance. In the image-to-video generation setting, it is hard for the attacker to robustly train a model from scratch due to the substantial computational resources and high-quality data required. Therefore, we focus on studying the impact of model fine-tuning on images protected by DORMANT. Specifically, we fine-tune the models of Animate Anyone using 1,000 protected images (with corresponding original images as targets) and 1,000 clean images from the TikTok dataset, yielding FID-FVD, FVD, LPIPS, FID, PSNR, and SSIM values of 93.52, 810.11, 0.432, 109.23, 15.01, and 0.615, respectively. While robust fine-tuning reduces DORMANT's effectiveness to some extent, the protective effect remains sufficient to ensure that the generated videos maintain low quality.

Robustness against Targeted Removal. We assume that the attacker has fully white-box access to DORMANT's mechanism, including the same pre-trained models and exact objective functions. Inspired by [3] which optimizes the purified image by pushing it in the opposite direction of the protected image and pulling it in the same direction of a visual reference, we implement this strategy in our proposed $\mathcal{L}_{\text{DORMANT}}$ and use purified images generated by DiffPure as visual references during optimization. The resulting FID-FVD, FVD, LPIPS, FID, PSNR, and SSIM values are 52.41, 485.01, 0.378, 97.73, 17.66, and 0.706, respectively. When compared to DiffPure's results presented in Table 4, we find that white-box access does not lead to a noteworthy increase in the purification effect. While DiffPure remains the most effective among the evaluated purifications, future work could explore ways to further improve DORMANT's robustness against such diffusion-based adversarial purification methods.

C More Ablation Study

Impact of Pose Repeats F and Decay Factor μ. F defines the length of the pose sequence used in $\mathcal{L}_{frame}$; μ controls the influence of previous gradients during each PGD step. As shown in Figure 14, both the increase in F and μ cause the protection effect to first increase and then decline. While we optimize the protective perturbation to induce frame incoherence by maximizing the distance between each generated frame, increasing F leads to more pairs of distances to compute, making the optimization more difficult and also increasing GPU memory usage. For the decay factor μ, moderate memorization of previous gradients helps stabilize update directions and avoid poor local maxima, but excessive impact from previous gradients with a large μ would hinder the search for new update directions. We set the default values of F and μ to 5 and 0.5, both within the sweet spot region.



(a) Pose repeats F (b) Decay factor μ

Figure 14: Impact of pose repeats F and decay factor μ.

D Visualized Results of Anomalies

As discussed in Section 5, we find that Glaze outperforms DORMANT on sample 002, and PhotoGuard and SDS perform better on sample 006 of the TikTok dataset. Here we present the visualizations of generated videos in Figure 15. While DORMANT has demonstrated sufficient effectiveness in these cases, these baseline protections perform even better on certain samples. Figure 16 visualizes generated video frames against commercial services Gen-2 and Gen-3, where DORMANT performs relatively poorly, mainly inducing identity mismatches without significant distortion.

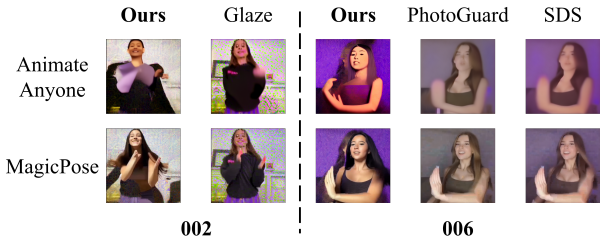


Figure 15: Visualizations of cases where DORMANT does not outperform baseline protections.

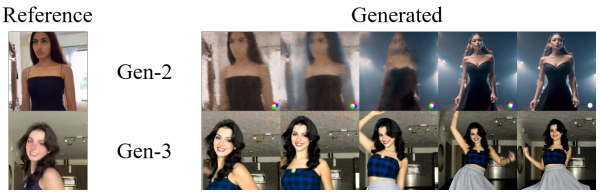


Figure 16: Visualizations of cases where DORMANT performs relatively poorly.

E Image Similarity

We evaluate the similarity between the protected images and their corresponding original images, and present the PSNR for different protection methods in Figure 17. AntiDB, Mistv2, and VGMShield offer better invisibility but with limited effectiveness in Table 1; Glaze, PhotoGuard, and DORMANT exhibit comparable invisibility; and SDS shows the lowest invisibility. While these similarities are computed under the perturbation budget η of $16/255$, users can customize η to strike a balance between invisibility and protection effectiveness based on their needs, as mentioned in Section 4.6.

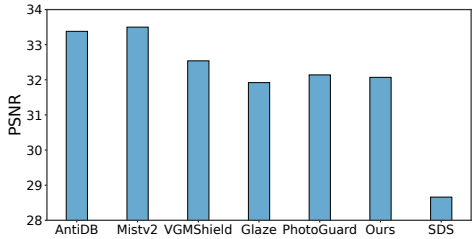


Figure 17: Image similarity before and after protections.

F Details of Human and GPT-4o studies

F.1 Questionnaire in Human Study

We detail our questionnaire used in the human study, which is divided into two parts: demographics and video selection, consisting of a total of 15 single-choice questions.

Part I - Demographics.

Q1: How old are you? [18-24], [25-34], [35-44], [45-54], [55-64], [65+], [Prefer not to say]

Q2: What gender do you best identify with? [Male], [Female], [Non-binary], [Prefer not to say]

Q3: What is the highest level of education you have completed or are currently pursuing? [High school], [Bachelor's], [Master's], [PhD], [Other], [Prefer not to say]

Q4: What is your field of expertise? [Computer Science], [Engineering], [Business/Finance], [Arts/Humanities], [Social Sciences], [Health Sciences], [Natural Sciences], [Other], [Prefer not to say]

Q5: How familiar are you with the technique of image-to-video generation? [Not familiar at all - I have never heard of it], [Not very familiar - I have heard of it but don't know much about it], [Somewhat familiar - I know the basics], [Familiar - I have a good understanding], [Very familiar - I am highly knowledgeable], [Prefer not to say]

Part II - Video Selection.

Q6 - Q15: Please select the video (A, B, C, D, or E) that you believe shows the most significant differences in facial features compared to the reference video (labeled 'REF' in the top left corner). Your choice should prioritize the video that best ensures the protection of portrait rights and privacy. (*Note: These videos are in GIF format and can be played by clicking.*) [A], [B], [C], [D], [E]

Figure 18: Example of the video selection question.

F.2 Prompt in GPT-4o Study

We detail our query prompt used in the GPT-4o study. The study is conducted on ChatGPT Web via a browser, using a ChatGPT Plus account. Frame sequences are converted into image grids in PNG format and uploaded via "Attach files".

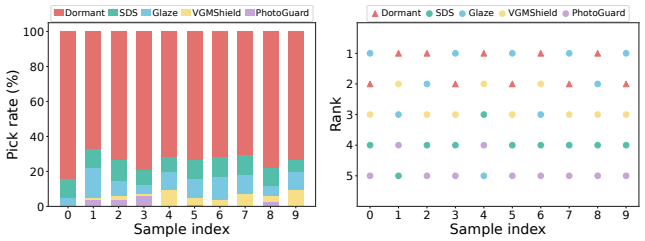
Query Prompt to GPT-4o

You will be given six frame sequences: A, B, C, D, E, and a reference frame sequence (REF). Your goal is to rank the quality of the five frame sequences (A, B, C, D, E) from most to least different compared to the reference frame sequence (REF). Base your ranking on how effectively each frame sequence protects the portrait and privacy rights of the individual.
To evaluate the effectiveness of portrait and privacy protection, please consider factors that differentiate each frame sequence from the reference, such as mismatches in the appearance of the human identity, changes in the background, and any additional observable signs. Your evaluation should consider, but not be limited to, these list factors provided merely as examples. Try to explore and assess as many other potential factors as possible.
Please analyze step by step and output your ranking for the five sequences, along with the reasoning which should detail how you evaluate each sequence. You should avoid any potential bias in your evaluation, and ensure that the order in which the frame sequences are presented does not affect your judgment.

Output Format:
Rank: [ordered list from most to least effective]
Reason: [step-by-step analysis]

F.3 Detailed Results

Figure 19 presents detailed results for each question in the human and GPT-4o studies. In the human study, DORMANT achieves a pick rate of at least 67% across all samples, with a maximum of 84.15%. Similarly, DORMANT is consistently ranked first or second across all samples by GPT-4o. These results highlight the superiority of DORMANT over baseline protections from the perspectives of both human perception and the advanced multimodal large language model.



(a) Results of human study (b) Results of GPT-4o study

Figure 19: Detailed results of human and GPT-4o studies.