Quality-Preserving Imperceptible Adversarial Attack on Skeleton-based Human Action Recognition

Ziyi Chang, Kanglei Zhou, Xiaohui Liang, Hubert P. H. Shum, *Senior Member, IEEE*

***Abstract*—Adversarial attacks on skeletal human action recognition have received significant attention. However, existing methods typically introduce noise-like perturbations that degrade motion quality post-attack, and thereby are inherently perceptible with recent advancements in S-HAR systems. We discover that this degradation stems from the gap between empirical and true risks during the optimization process of previous adversarial attacks. To address this issue, we propose an attack where adversarial motions are obtained without compromising their motion quality. To minimize the risk gap and preserve motion quality, we propose a distribution-based adversarial attack method without introducing noise-like perturbations. To faithfully evaluate the motion quality, we propose a new metric that aligns with human perception on real-world naturalness. Experiments have been conducted on the state-of-the-art S-HAR methods across two datasets, demonstrating the superiority of our method in both the attack success rate and the post-attack motion quality through qualitative and quantitative analyses. The success of our quality-preserving attack application and distribution-based method raises serious concerns about the robustness of action recognizers, highlighting the need for further enhancements in this domain.**

***Index Terms*—Adversarial Attack, Human Action Recognition, Human Skeletal Motion, Diffusion Model, Motion Quality.**

I. Introduction

ADVERSARIAL attacks on skeletal human action recognition (S-HAR) have received significant attention due to concerns regarding the robustness of S-HAR systems. These systems have been deployed to recognize human actions from skeletal inputs and have been deployed in various security-critical and human-centric domains such as medical care, action assessment, and safety surveillance [1]. The adversarial motions pose substantial threats to such life-critical applications by misleading the S-HAR systems [2]. Therefore, obtaining adversarial motions is important to enhance the robustness of these S-HAR systems and boost trustworthy and responsible artificial intelligence [3], [4].

Previous adversarial attack methods [5], [6] typically introduce noise-like perturbations to skeletal inputs to deceive S-HAR algorithms, resulting in reduced post-attack motion quality. These perturbations inherently conflict with human perceptions of natural motion [7], [8] due to the quality decline, and thereby diminish the imperceptibility of existing attacks. Nonetheless, this vulnerability is often overlooked due to inherent dataset noise [9], [10], [11] and limited classifier expressiveness [12], [13], [14]. Furthermore, existing metrics that compare paired pre-attack and post-attack motions inadequately assess post-attack motion quality because motions are sparsely distributed and their neighborhoods may not be natural and plausible [15]. With advancements enabling high-quality skeletal inputs for S-HAR systems [16], [17], [18], the attacked motions become increasingly perceptible and undermine the imperceptibility of adversarial attack.

We identify the decline in the quality of adversarial motions resulting from the optimization process in previous attack methods, wherein the empirical risk for optimization significantly deviates the true risk. The empirical risk typically refers to the loss that is defined over a large number of observations while the true risk is defined as the expected loss across the implicit data distribution [19]. However, in the adversarial attack context, only an individual motion is involved for optimizing a motion to be adversarial, which leads to a large gap between the empirical and true risks [20]. With the large risk gap, optimization with respect to a single motion observation commonly introduces the noise-like patterns [21], decreasing the quality of adversarial motions.

To tackle the decline in the quality of adversarial motions, we propose a novel attack application where adversarial motions are obtained without sacrificing quality. This application demonstrates a new task towards imperceivable adversarial attack where the post-attack motion quality are explicitly constrained rather than the paired-motion constraints in previous methods. To minimize the risk gap in the optimization process of adversarial attack, we propose a distribution-based S-HAR attack method where a diffusion model is employed to learn the data distribution. The proposed distribution-based method relies on the data distribution instead of paired pre-attack and post-attack motions to optimize the given motions. By reducing the risk gap, our optimization does not introduce noise-like perturbations and thereby our adversarial motions are obtained without sacrificing the motion quality. Furthermore, to faithfully quantify the quality of adversarial motions, we propose a new metric to measure the naturalness based on physiological analysis of human movements [22], [23]. This new metric measures the inherent physiological naturalness rather than rely on the paired individuals comparison in existing adversarial metrics, aligning with the human perceptions

Manuscript received February 6, 2025. This research is supported in part by the EPSRC NorthFutures project (ref: EP/X031012/1). (*Corresponding author: Hubert P. H. Shum.*)

Ziyi Chang and Hubert P. H. Shum are with the Department of Computer Science, Durham University, Durham, DH1 4FL, UK (e-mail: {ziyi.chang, hubert.shum}@durham.ac.uk).

Kanglei Zhou is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: zhoukanglei@buaa.edu.cn).

Xiaohui Liang is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Zhongguancun Laboratory, Beijing 100190, China (e-mail: liang_xiaohui@buaa.edu.cn).

of natural motions in the real world.

Extensive experiments show that our method generates adversarial motions with the least quality decrease for attacking against the four latest S-HAR classifiers across both a high quality dataset, i.e., 100STYLE [24] and a commonly used dataset, i.e., HDM05 [25] when compared with state-of-the-art S-HAR adversarial attack methods. Our user study further validates that our adversarial motions are the least perceivable by humans. To further validate our method, we conduct ablation studies on hyper-parameters of diffusion models and the specification of generative models. Codes are available in https://github.com/mrzzy2021/QualityPreservingAttack. Our contributions are summarized as follows:

- We discover a critical yet previously overlooked vulnerability in existing attack methods where their noise-like adversarial perturbations are inherently perceptible to humans. To address this, we propose a novel attack application that imperceptible adversarial attack is fulfilled without sacrificing post-attack motion quality.
- To preserve post-attack motion quality, we propose a distribution-based attack method where imperceptible adversarial motions are generated through a pre-trained diffusion model by minimizing the gap between empirical and true risks in our optimization of adversarial attack.
- To faithfully quantify the quality of adversarial motions, we propose a new metric based on existing physiological analysis of real-world human motions, enabling the assessment of previously overlooked vulnerabilities that existing metrics fail to capture.

The remaining of this paper is organized as follows: We first review related work in Section II. Then, we formulate the proposed quality-preserving adversarial attack application in Section III. Subsequently, we present our proposed distribution-based adversarial attack method with quality preservation in Section IV. We formulate the proposed metric in Section V to faithfully evaluate the motion quality after attack. Finally, we demonstrate the superiority of our method in Section VI and provide a summary in Section VII with potential future work.

II. Related Work

A. General Adversarial Attack

Adversarial attacks are first introduced in [26], highlighting the vulnerability of deep neural networks and subsequently extending to various data types. Generally, adversarial attacks serve as a specialized form of data augmentation aimed at exposing system vulnerabilities by generating new samples, whereas other data augmentation techniques may focus on objectives such as enhancing training efficiency and improving inference performance [27]. Deep neural networks remain susceptible to meticulously crafted adversarial attacks despite that significant achievements have been witnessed. Noise-like perturbations have been applied to input data to easily deceive these high-performing neural networks and people raise concerns on the trustworthiness and reliability of deployed neural networks [3], [4]. In response to these concerns, researchers have extensively explored adversarial attacks across different data modalities, including 2D images [28], videos [29], 3D objects [30], physical objects [31], and graph data [32]. While adversarial research on different modalities has been widely investigated, attacks on time-series data have recently gained attention [33], with relatively minimal focus on 3D skeletal motions characterized by complex spatio-temporal structures.

B. Adversarial Attacks on S-HAR

Adversarial attacks on skeletal human action recognition (S-HAR) aim to perturb 3D human skeletons to deceive classifiers, mirroring objectives from the image domain. Skeletal motion is extensively utilized in HAR to mitigate challenges such as lighting variations, occlusions, and diverse view angles [1]. Consequently, the vulnerability of skeleton-based classifiers to adversarial attacks has received increasing attention. With significantly lower degree of freedom in motions [34], [35], previous S-HAR attack methods focus on improving better imperceptibility through different perspectives. [6] targets GCN-based models by employing generative adversarial networks as a discriminator to achieve imperceptibility in terms of the anthropomorphic plausibility. [36] proposes a new attack method where only the length of bones could be perturbed to enhance the imperceptibility from the perspective of bone length consistency. [5] analyzes the perceptibility of adversarial skeletal samples and proposes constraints between paired pre-attack and post-attack motions, integrating the observations from [6] and [36] to further improve imperceptibility. However, while these methods effectively achieve to fool the S-HAR system, they inherently introduce noise-like perturbations that degrade post-attack motion quality, and thereby undermine the imperceptibility of post-attack motions.

C. Diffusion Models

Diffusion models [37] have attracted extensive attention for their powerful generative capabilities. These models utilize numerous timesteps to transform data distributions into a standard Gaussian distribution through a forward process, while a denoising network is trained to recover the less noisy data distribution [38]. A large number of timesteps are typically required for smooth distribution transitions and simultaneously provide a hierarchical latent space with rich semantics.

Some have leveraged the generative ability of diffusion models for adversarial attack. [39] directly learns adversarial generation using paired adversarial samples, though this dependency may not be practical for real-world applications. [40] introduces a label-based generative method, training diffusion models on auxiliary datasets. [41] leverages diffusion models with modified denoising process to generate adversarial samples. [28] utilizes diffusion models for high-quality image steganography. [42] focuses on purification as a defense approach by the generative ability of diffusion models rather than the adversarial attack. Despite the applications mentioned above, the potential of leveraging the rich semantics extracted in the hierarchical latent space of diffusion models has been overlooked in the adversarial attack field, leading to complicated training re-design or meticulously crafted datasets.

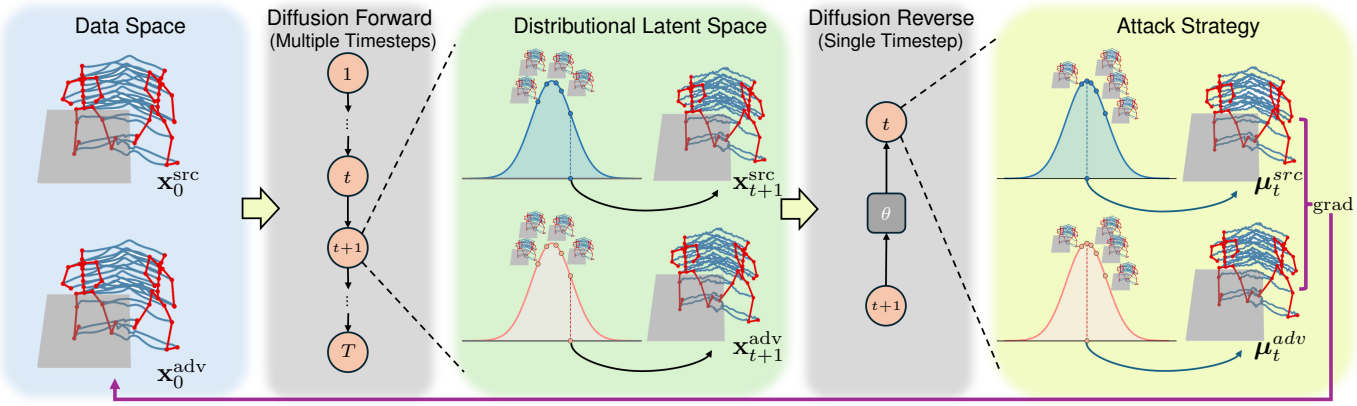


Fig. 1: Overview of the proposed distribution-based imperceivable adversarial S-HAR attack where post-attack motion quality is preserved. We optimize the motions along the gradient provided by a pretrained diffusion model based on the distributional latent space to keep the post-attack motion quality.

III. Imperceptibility of S-HAR Adversarial Attack via Motion Quality Preservation

In this section, we formulate our proposed attack application, which conducts imperceptible adversarial attack against skeleton-based human action recognition (S-HAR) systems without compromising post-attack motion quality.

a) Inherent Vulnerability of Imperceptibility: Existing adversarial attacks achieve imperceptibility by introducing small, noise-like perturbations that inherently and detrimentally affect post-attack motion quality. Motion quality encompasses naturalness and plausibility [43], as human motions adhere to physical and biomechanical constraints. The human brain has specialized neural mechanisms for perceiving biological motion [44], [45] and is highly sensitive to unnatural and implausible kinematics [7], [8]. Given the highly nonlinear and articulated spatial-temporal structures of human motions [5], even subtle noise-like perturbations significantly degrade quality. Such perturbations disrupt motion dynamics and physical constraints, and deviating motions from the natural and plausible distribution. Thus, perturbed motions suffer from the decline in motion quality and become perceptible.

b) Task Formulation: Nonetheless, the inherent perceptible vulnerability resulting from the introduced noise-like perturbations is often overlooked by previous S-HAR attack methods due to the intrinsic noise in datasets [9], [10], [11] and the limited expressiveness capacity of classifiers [12], [13], [14]. As advancements [16], [17], [18] have widely facilitated cutting-edge S-HAR systems to utilize skeletons with high quality as inputs for decision-making, human supervisors readily perceive suspicious noise-perturbed skeletal inputs.

Furthermore, previous constraints during their optimization process, which usually focus on paired pre-attack and post-attack motions, fail to faithfully reflect the quality of adversarial motions. Since motions are sparsely distributed [15], attacked motions that remain within the neighborhood of their pre-attack motions usually may not be natural and plausible, especially when considering the motion dynamics. Therefore, adversarial motions suffering from post-attack quality becomes increasingly perceptible due to the added noise-like perturbations, which urgently needs to be resolved to improve the imperceptibility of adversarial attacks.

To achieve imperceptibility by preserving the quality of adversarial samples instead of introducing noise-like perturbations, we propose an adversarial task, in which post-attack motions preserve pre-attack motion quality. Specifically, our task explicitly requires the quality of the post-attack motions as the optimization constraint instead of individual motion pairs, and aims to achieve imperceptibility by preserving post-attack motion quality. We define our adversarial task as follows:

$$\begin{aligned} \underset{\mathbf{x}^{\text{adv}}}{\text{minimize}} \quad & \mathcal{L}_{\text{adv}}(C(\mathbf{x}^{\text{adv}})) \\ \text{subject to} \quad & \mathcal{Q}(\mathbf{x}^{\text{adv}}) = \mathcal{Q}(\mathbf{X}^{\text{src}}), \end{aligned} \tag{1}$$

where $C(\cdot)$ is an S-HAR classifier, $\mathbf{x}^{\text{adv}} \in \mathbb{R}^{T\times J\times 3}$ represents the attacked motion, $\mathbf{X}^{\text{src}} \in \mathbb{R}^{N\times T\times J\times 3}$ represents the set of natural and plausible motions, and $Q(\cdot)$ represents the motion quality. Compared with previous adversarial attacks, we propose optimizing an individual motion sample, $\mathbf{x}^{\text{adv}}$, with respect to the entire set of natural and plausible motions rather than rely on a pair of pre-attack and post-attack motions.

IV. Distribution-based S-HAR Attack Method

In this section, we demonstrate the proposed distribution-based adversarial attack method against skeleton-based human action recognition (S-HAR) systems. First, we identify the origin of noise-like perturbations introduced in existing attack methods and propose integrating the data distribution into adversarial attacks by constructing a generative diffusion distributional latent, as described in Section IV-A. Then, we present our attack strategy to leverage the constructed latent for a quality-preserving adversarial attack in Section IV-B.

A. The Diffusion Latent for Quality Preservation

a) The Risk Gap of Previous Optimization: The noise-like perturbations in previous attacks are attributed to the gap between empirical and true risks in their optimization processes. Empirical risk typically refers to the loss defined over a large number of observations, while true risk is defined as the expected loss across the implicit data distribution [19]. However, in previous adversarial attacks, the optimization process involves only an individual motion sample, leading

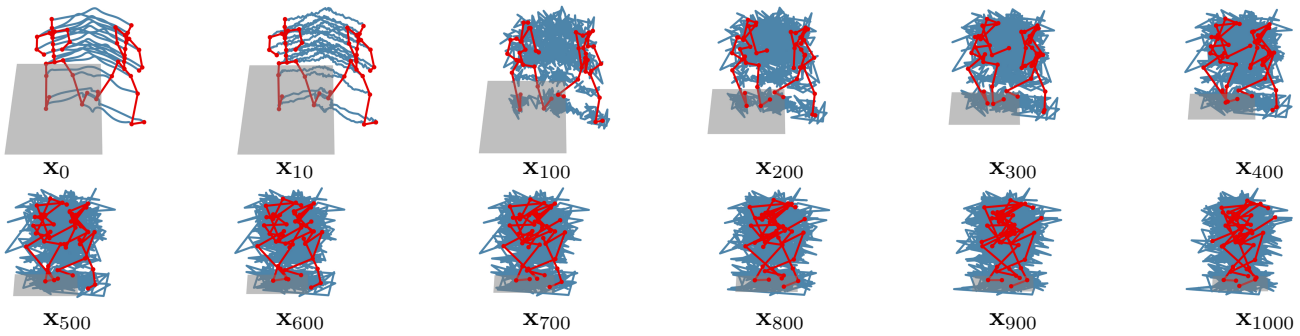


Fig. 2: The visualization of diffusion latents at different timesteps. As shown, the earlier timesteps maintain more low-level details, the later timesteps focus on high-level structures until latents become pure noise.

to a large gap between empirical and true risks [20]. When optimizing a motion with respect to a single pre-attack motion observation, noise-like perturbations are usually introduced [21] and decrease the quality of adversarial motions.

b) Constructing Generative Diffusion Latent: To minimize the risk gap in the optimization process of adversarial attacks, we propose leveraging distributional latents from generative diffusion models for our quality-maintaining adversarial attack. As shown in Fig. 2, diffusion models have semantically meaningful latent spaces that capture different feature patterns with respect to various choices of timesteps [46]. When compared to discriminative models [47], they are capable of modelling motion distributions that are critical for quality preservation. Compared with other generative models like VAEs [48], their hierarchical structures are more comprehensive to represent underlying patterns to modify motions.

Our attack leverages diffusion distributional latents as proxies for the given pre-attack motion individual. Previous S-HAR adversarial attack methods typically optimize motions directly in the original data space. However, the original space is sparse where natural motions being sparsely distributed [15]. This sparsity leads to difficulties in modifying motions within the original data space while maintaining their quality. In contrast, we build our attack method on the stochastic latent space of diffusion models. The stochastic distributional latents are located in am approximately smooth space [49] and encode different levels of semantics for adversarial modification. By transforming the given pre-attack motion into latent distributions, the empirical risk of our optimization involves infinite samples within distributions rather than individual motions.

Specifically, we define our latent for adversarial attack to be the posterior mean due to its rich semantics encoded in the stochastic process of diffusion models. The posterior mean combines both the forward and reverse processes in a diffusion model, thereby comprehensively and hierarchically representing the motion. First, we leverage the forward process to convert motions into a distribution, which is defined as:

$$p(\mathbf{x}_{t+1}|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{t+1}}\mathbf{x}_0, (1-\alpha_{t+1})\mathbf{I}), \tag{2}$$

where $\mathbf{x}_0$ is a pre-attack motion, and $\mathbf{x}_{t+1}$ is a stochastic sample at the timestep $t+1$ from a Gaussian distribution. Then, to achieve finer-grained latent manipulation for adversarial attack, we propose to construct our latent proxy for adversarial attack in a single timestep of diffusion reverse process. The posterior distribution in the reverse process is defined as:

$$p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0) = \mathcal{N}(\boldsymbol{\mu}, \sigma_t\mathbf{I}), \tag{3}$$

where $\mathbf{x}_{t+1} \sim p(\mathbf{x}_{t+1}|\mathbf{x}_0)$, σ_t is the posterior variance and $\boldsymbol{\mu}$ is the posterior mean of the distribution. Since the posterior variance is pre-defined in noise schedule, $\boldsymbol{\mu}$ represents the denoising direction according to the class label. To show relationship between the posterior mean and data distribution, we explicitly derive the posterior mean based on Tweedie's formula [50] and formulate it as:

$$\boldsymbol{\mu}_t = \gamma_t\mathbf{x}_0 + \lambda_t\epsilon_t + \delta_t\nabla \log p_\theta(\mathbf{x}_{t+1}, \mathbf{y}), \tag{4}$$

where $\mathbf{x}_0$ is the motion in original data space, $\mathbf{x}_{t+1}$ is sampled from the latent distribution, i.e., $\mathbf{x}_{t+1} \sim p(\mathbf{x}_{t+1}|\mathbf{x}_0)$ and $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. The posterior mean is a latent that encapsulates the prior knowledge of the entire dataset's distribution. The first term, $\mathbf{x}_0$ ensures content preservation during the attack, while the distributional term is responsible for aligning the data distribution to remain post-attack motion quality.

c) Benefits of Distributional Latent: As shown in Eq. 4, the constructed latent is the posterior mean derived from the data distribution rather than from a single sample. Compared with deterministic latents [51], the stochastic latents are more efficient and have much higher capacity [46]. Moreover, the posterior mean represents a direction that conforms to the data distribution toward high-density areas associated with the given labels. When we optimize over this latent for the adversarial attack, our empirical risk is based on all data samples from the approximated distribution, instead of a single sample. Additionally, instead of computing over the entire diffusion chain, we focus on only one denoising step to minimize changes in a single optimization iteration. Through sampling different timesteps, the constructed latent captures different features that encode underlying semantics.

B. The Attack Strategy for Quality-Preserving Attack

a) Overview of Strategy: To leverage the diffusion posterior latent space for adversarial attacks, we propose perturbing the source motion observation by randomly sampling different class labels rather than relying on the gradient of a specific classifier. This is not only because obtaining the gradient of a

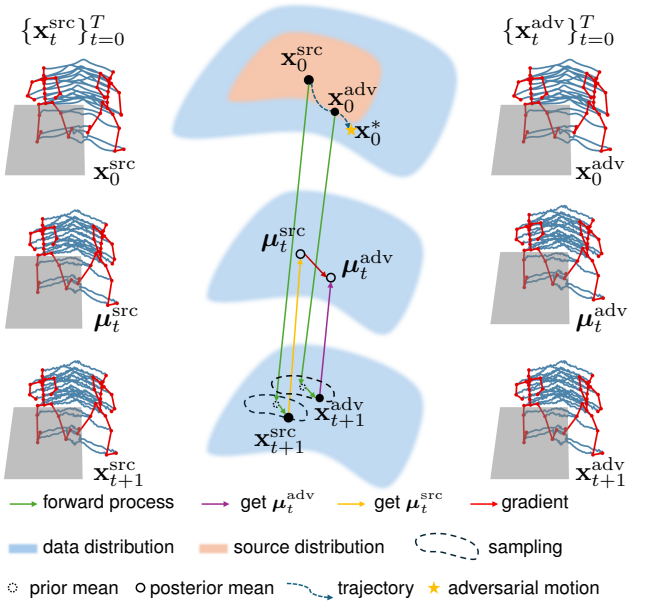


Fig. 3: The illustration of attack strategy. We illustrate an intermediate calculation at the timestep t during the optimization of achieving the final adversarial motion $\mathbf{x}_0^*$.

specific classifier is challenging in real-world applications [2], but also because the gradient of a classifier may not be reliable for preserving motion quality. As classifiers focus on the label distribution rather than data distribution, their gradient may point to out-of-distribution regions. While following this gradient achieves the shortest trajectory of deceiving classifiers, it leads to a decline in post-attack motion quality and undermines the imperceptibility. Since conditional diffusion models themselves serve as generative classifiers [52], they provide directions to class boundaries that conform to data distributions. As a result, we design our method to only rely on the classifier's decisions and drive the motion towards adversarial samples by sampling adversarial labels for the conditional diffusion model.

b) Optimization Objective: Specifically, we denote the stochastic distributional latents of the source motion observation and the adversarial motion as $\boldsymbol{\mu}_t^{\text{src}}$ and $\boldsymbol{\mu}_t^{\text{adv}}$. The source motion observation and the adversarial motion are mapped to the latent space by the diffusion forward process and then the desired posterior mean are obtained within a single timestep denoising conditioned on the ground truth label $\mathbf{y}^{\text{src}}$ and the randomly sampled the adversarial label $\mathbf{y}^{\text{adv}}$ from the set of all possible labels excluding the ground truth label, respectively. We define our objective function as follows:

$$\mathcal{L}_{\mu_t} := 0.5 \times \mathbb{E}_{t,\epsilon_t}\left[\|\boldsymbol{\mu}_t^{\text{adv}} - \boldsymbol{\mu}_t^{\text{src}}\|_2^2\right], \tag{5}$$

where $\boldsymbol{\mu}_t^{\text{src}}$ and $\boldsymbol{\mu}_t^{\text{adv}}$ are our defined latents and serve as proxies for adversarial attack. The first term, $\boldsymbol{\mu}_0^{\text{adv}}$, represents the direction towards being adversarial, while the second term, $\boldsymbol{\mu}_0^{\text{adv}}$, represents the direction of maintaining the representative semantics within the original class. Our optimization objective is an expectation over the diffusion timesteps t and the stochastic distributional latent, indicated by the randomly sampled noise ϵ in the forward process where semantics are encoded into the latent space of the diffusion model.

Algorithm 1 Diffusion-based Quality-Preserving Adversarial Motion Attack on S-HAR

Require: Diffusion model θ, a classifier φ, a motion $\mathbf{x}_0^{\text{src}}$ with label $\mathbf{y} \in \mathbf{Y}$, maximum iteration I, diffusion timesteps T

$\mathbf{x}_0^{\text{adv}} \leftarrow \mathbf{x}_0^{\text{src}}$
$i \leftarrow 0$
$\mathbf{y}^{\text{adv}} \sim \mathbf{Y}_{/\,\mathbf{y}^{\text{src}}}$ ▷ Randomly sample an adversarial label
while $i \leq I$ *and* $\mathbf{y}^{\text{pred}} \neq \mathbf{y}^{\text{src}}$ **do**
  $t \sim [1, T]$
  $\mu_t^{\text{src}} = \mu_t(\mathbf{x}_0^{\text{src}}, \mathbf{y}^{\text{src}}; \theta)$ ▷ Eq. 4
  $\mu_t^{\text{adv}} = \mu_t(\mathbf{x}_0^{\text{adv}}, \mathbf{y}^{\text{adv}}; \theta)$ ▷ Eq. 4
  $\text{grad} = \mu_t^{\text{adv}} - \mu_t^{\text{src}}$ ▷ Eq. 6
  $\mathbf{x}^{\text{adv}} = \mathbf{x}^{\text{adv}} + \text{grad}$
  $\mathbf{y}^{\text{pred}} = \arg\max p_\varphi(\mathbf{y}|\mathbf{x}^{\text{adv}})$ ▷ Get classifier decision
end while

By minimizing the defined object, our optimization aligns the stochastic distributional latents of the source and the adversarial motions. Specifically, we calculate the gradient of $\mathcal{L}_{\mu_t}$ with respect to $\boldsymbol{\mu}_t^{\text{adv}}$ and obtain our adversarial gradient:

$$\begin{aligned} \text{grad} :=& \nabla \mathcal{L}_{\mu_t} \\ =& \mathbb{E}_{t,\epsilon_t}\left[\mu_t^{\text{adv}} - \mu_t^{\text{src}}\right], \end{aligned} \tag{6}$$

on which we rely to iteratively update the $\mathbf{x}_0^{\text{adv}}$. The Eq. 6 facilitates the adversarial effect and quality preservation. Instead of directly requiring the paired pre-attack and post-attack motions to be close with each other in data space, our optimization aligns the posterior distribution of pre-attack and post-attack motions. As the noise schedule is predefined, the posterior distribution, which is a Gaussian distribution, is controlled by the posterior mean. Constraining the two posterior means $\boldsymbol{\mu}_t^{\text{adv}}$ with $\boldsymbol{\mu}_t^{\text{src}}$ ensures that the posterior distributions of $\mathbf{x}_0^{\text{adv}}$ and $\mathbf{x}_0^{\text{src}}$ remain closely aligned rather than merely examine the pre-attack and the post-attack motions. Our optimization represents a single timestep examination where a multi-timestep generative process conditioned on $\mathbf{y}^{\text{adv}}$ is usually employed for generation. By optimizing over the expectation of timesteps, the trajectory defined by the posteriors is expected to be closely aligned with each other. Consequently, our attack strategy facilitates the generation of $\mathbf{x}_0^{\text{adv}}$ that corresponds with $\mathbf{y}^{\text{adv}}$ to deceive the target model, while simultaneously preserving the quality of $\mathbf{x}_0^{\text{src}}$. Detailed information regarding our adversarial attack methodology is provided in Algorithm 1 and Fig. 3.

c) Relationship with Previous Optimization: We further demonstrate that our proposed method implicitly integrates previous approaches while additionally offer the advantage of distributional prior knowledge provided by a pre-trained diffusion model to minimize the risk gap. The Eq. 6 is equivalently represented using the input motion and the learned distribution,

from which the following detailed formulation is derived:

$$\begin{aligned}\text{grad} &:= \mathbb{E}_{t,\epsilon_t}\left[\mu_t^{\text{adv}} - \mu_t^{\text{src}}\right] \\ &= \mathbb{E}_{t,\epsilon_t}\Bigg[\psi(t)(\underbrace{\mathbf{x}_0^{\text{adv}} - \mathbf{x}_0^{\text{src}}}_{\substack{\text{Distance Function} \\ \text{in Original Space}}}) \\ &\quad + \chi(t)\Bigg(\underbrace{(\nabla \log p(\mathbf{x}_t^{\text{adv}}) - \nabla \log p(\mathbf{x}_t^{\text{src}}))}_{\text{Distribution Constraint}} \\ &\quad + \underbrace{(\nabla \log p(\mathbf{y}^{\text{adv}}|\mathbf{x}_t^{\text{adv}})}_{\substack{\text{Adversarial} \\ \text{Gradient}}} - \underbrace{\nabla \log p(\mathbf{y}^{\text{src}}|\mathbf{x}_t^{\text{src}}))}_{\substack{\text{Representiveness} \\ \text{of the Given Input}}}\Bigg)\Bigg],\end{aligned} \tag{7}$$

where we replace the $\boldsymbol{\mu}_0$ with Eq. 4 and then decompose the probability with Bayes' theorem. The four terms serve distinct yet complementary purposes in our objective function. The first term demonstrates a similar measurement to previous attack methods, ensuring that the adversarial sample does not deviate excessively from the input sample. However, our ability of quality preservation, which distinguish our method from previous methods, is ensured through the remaining terms. Unlike traditional adversarial optimization, which considers only an individual motion pair as shown in the first term, our perturbation strategy provides the modification gradient based on data distributions that are composed of infinite motions. The second term measures the distribution density and ensures that the adversarial motions remain not only neighborhood but also high density area. The third term considers the adversarial gradient while the fourth term involves the representativeness of the observed sample as a consideration for the range of pre-attack motion neighborhood by evaluating how representative the given input is with respective to the label. These considerations leverage the distributional prior knowledge and enables the quality preservation under adversarial modifications.

V. Human Perception Aligned Naturalness Metric

In this section, we examine previous metrics and formulate the proposed metric to faithfully measure the quality of generated adversarial motions in terms of naturalness.

a) Unfaithfulness of Existing Metrics: The evaluation of existing metrics fails to accurately measure quality. Previous metrics rely on paired comparisons between the pre-attack and post-attack motions to assess quality, especially naturalness [5]. However, remaining within the neighborhood of a pre-attack motion does not ensure motion quality comparable to clean motions because natural and plausible motions are sparsely distributed [15]. Consequently, these metrics cannot reliably determine whether adversarial motions are sufficiently natural due to their misalignment with human perception.

b) Human Perception Aligned Metric: Natural human motions are subject to biomechanical constraints where movements are driven by muscle activation. Every action results from the brain sending electrical signals to nerves, which in turn contract muscles to move joints [53]. The human brain develops prior knowledge of natural motions based on real-world observations of biomechanically validated musculoskeletal movements [54] rather than paired comparison. Therefore, human perception is inherently biased toward the naturalness self-embedded in the observed motion.

To faithfully reflect naturalness, we propose a novel metric that evaluates the physiological naturalness of generated adversarial motions. Since real-world movements are created by muscle activation, our metric is grounded in the bio-activity of muscles. Specifically, the muscle activation patterns of natural human motions are constrained by physiological characteristics [22], [23], resulting in smooth activation curves. We introduce a new metric to measure quality from the biomechanical perspective of muscle activation:

$$\text{naturalness} = \frac{1}{T}\int_0^T \frac{\mathrm{d}^4 J(t)}{\mathrm{d}t^4}\mathrm{d}t, \tag{8}$$

where T is the temporal length of a motion, $J(t)$ is the joint position at time t.

VI. Experiments

In this section, we begin by outlining the experimental settings. Subsequently, we quantitatively and qualitatively analyze the performance of our proposed method and existing adversarial techniques. Finally, we perform a user study to empirically evaluate the imperceptibility of post-attack motions and also ablation studies to validate our configurations.

A. Experimental Settings

a) Datasets: To evaluate our imperceptible adversarial attacks via quality preserving, we select the 100STYLE [24] dataset due to its noise-free and inherently high-quality characteristics. The 100STYLE dataset is collected using a motion capture system and comprises 100 classes of different styles. We represent the skeletons using Cartesian coordinates for 23 joints. This dataset is pre-processed by segmenting long sequences into several segments according to the valid periods provided by [24]. Additionally, we employ the HDM05 [25] dataset to evaluate our method on a smaller scale dataset. Although HDM05 is commonly used for action recognition tasks, it has slightly lower quality compared to 100STYLE. We adhere to the pre-processing procedure outlined in [5]. The HDM05 dataset includes 65 classes of different human actions, and the hip joint is fixed to the origin after pre-processing.

b) Evaluated Models: Given the significant advancements in the field of human action recognition, we adopt the latest S-HAR models as victim classifiers to effectively evaluate the performance of adversarial attack methods against advanced classifiers. Specifically, we select the Style classifier [55], STTFormer [56], Skateformer [57], and FR-Head [58] as victim models. These models encompass both transformer-based and graph-based architectures. We utilize their publicly available codebases to train their models.

c) Evaluation Metrics: Human motions are governed by physical and biomechanical constraints, thereby necessitating the assessment of visual motion quality in terms of naturalness and plausibility [43]. To evaluate the motion quality in adversarial attacks, we introduce the physiological naturalness metric, as discussed in Section V, which is grounded in the physiological characteristics of natural human

TABLE I: Generated Adversarial Motion Quality Comparison on 100STYLE dataset [24].

Victim	Method	Success Rate ↑	FID ↓	MMD ↓	Physiological Naturalness ↓	Foot Skating ↓	Bone Variation ↓
Style [55]	I-FGSM	81.72%	194.95	0.053	129.42	0.147	10.96
	MI-FGSM	82.08%	195.22	0.053	131.33	0.147	11.16
	MIG	70.61%	238.23	0.071	264.46	0.234	23.22
	SMART	42.65%	242.12	0.072	97.73	0.119	4.00
	Ours	**100%**	**18.91**	**0.011**	**16.05**	**0.075**	**2.94**
STTFormer [56]	I-FGSM	80.29%	162.55	0.043	190.98	0.196	19.28
	MI-FGSM	80.29%	162.78	0.043	191.10	0.197	19.29
	MIG	72.76%	190.75	0.050	280.56	0.241	27.63
	SMART	29.39%	195.06	0.053	103.56	0.167	3.92
	Ours	**100%**	**20.28**	**0.013**	**16.42**	**0.077**	**3.09**
SkateFormer [57]	I-FGSM	68.46%	79.63	0.019	43.90	0.053	4.53
	MI-FGSM	68.82%	79.06	0.019	44.70	0.053	4.61
	MIG	60.22%	102.77	0.026	70.72	0.069	7.30
	SMART	31.90%	125.46	0.030	29.32	**0.047**	**1.60**
	Ours	**100%**	**20.16**	**0.014**	**16.33**	0.079	2.99
FR-Head [58]	I-FGSM	86.38%	226.80	0.068	312.90	0.276	25.65
	MI-FGSM	86.74%	226.67	0.068	316.77	0.280	26.08
	MIG	78.14%	242.85	0.075	489.37	0.396	38.12
	SMART	32.97%	262.75	0.100	214.63	0.301	6.33
	Ours	**100%**	**19.13**	**0.011**	**16.75**	**0.084**	**3.23**
Average	I-FGSM	79.21%	165.98	0.046	169.3	0.168	15.11
	MI-FGSM	79.48%	165.93	0.046	170.98	0.169	15.29
	MIG	69.34%	193.65	0.056	276.28	0.235	24.07
	SMART	34.23%	206.35	0.064	111.31	0.159	3.96
	Ours	**100%**	**19.62**	**0.012**	**16.39**	**0.079**	**3.06**

movements. Additionally, we report the Frechet Inception Distance (FID) and Maximum Mean Discrepancy (MMD) based on acceleration to measure the naturalness by assessing the distributional similarity between pre-attack and post-attack motions. As for plausibility, we report the foot skating ratio and bone length variations between frames. Following [59], foot skating is quantified by the consistency between foot velocity and foot height. A high foot skating ratio indicates significant violation in terms of physical constraints. Bone length variation [60] measures the consistency of bone lengths across frames. Higher deviations indicating distortions in the skeleton structure. Finally, we report the success rate of the adversarial attacks to evaluate the threatfulness.

d) Attacking Methods: We compare our method with the state-of-the-art (SOTA) S-HAR attack technique, i.e. SMART [5], as well as other adversarial attack methods including I-FGSM [61], MI-FGSM [62], and MIG [63]. To ensure a fair comparison, we execute 2000 iterations for each attack method, allowing all methods to explore a broader solution space in their pursuit of effective adversarial motions. Since our method requires a pre-trained diffusion model, we adopt the diffusion model proposed by [64] and follow the prescribed settings to pre-train it on the two datasets.

B. Adversarial Motion Quality Evaluation

We qualitatively and quantitatively evaluate the performance of adversarial atttack in terms of deceitfulness, motion quality including naturalness and plausibility, and human imperceptibility. Deceitfulness measures the effectiveness of an adversarial method. In addition to deceitfulness, motion quality assesses whether the adversarial motions are plausible and natural, and serves as an indicator for imperceptibility. Beyond analytical measurements, we empirically evaluate human imperceptibility. Imperceptibility demonstrates the possibility that humans cannot distinguish pre-attack and post-attack motions when they are mixed together as S-HAR inputs. For simplicity, we present only representative results in this paper.

a) Deceitfulness: The success rates presented in Table I indicate that our method is the most deceitful compared with other methods on the 100STYLE dataset. Our method consistently achieves an average success rate of 100%. Our attack effectively modifies nearly every input to become adversarial without relying on the gradient of victim model within a limited number of iterations. In contrast, other methods may struggle to generate adversarial motions within iteration constraints. This demonstrates that the stochastic latent features convey comprehensive semantics [65] about the underlying dynamics via different timesteps while the semantics cannot be easily represented in the original motion space.

b) Motion Quality: Regarding naturalness, the FID and MMD scores in Table I show that the distribution of our generated motions closely resembles that of natural motions, whereas other methods exhibit significant distributional deviations. Our adversarial motions maintain proximity to the ground truth distribution of motion dynamics by a considerable margin. The lowest FID and MMD scores indicate that our adversarial motions are indistinguishable from those in the ground truth dataset. In other words, the post-attack motion quality are well preserved because our method leverages the distributional knowledge introduced by the pre-trained diffusion model. By utilizing data distribution, our method minimizes the gap between empirical and true risks by assessing deviations between the modified motions and the data distribution. Conversely, other methods can only access single source motions, and disrupt the naturalness of post-attack motions, resulting in much higher FID and MMD scores.

TABLE II: Generated Adversarial Motion Quality Comparison on HDM05 dataset.

Victim	Method	Success Rate ↑	FID ↓	MMD ↓	Physiological Naturalness ↓	Bone Variation ↓
Style [55]	I-FGSM	93.91%	129.78	0.086	1226.33	126.39
	MI-FGSM	93.91%	129.79	0.086	1226.98	126.46
	MIG	92.47%	194.33	0.170	2273.09	238.34
	SMART	68.46%	142.39	0.116	1052.44	41.00
	Ours	**100%**	**22.91**	**0.012**	**187.28**	**32.15**
STTFormer [56]	I-FGSM	86.74%	162.98	0.124	1488.08	144.97
	MI-FGSM	86.74%	163.18	0.125	1487.92	144.95
	MIG	76.34%	202.24	0.188	1945.83	189.95
	SMART	75.63%	174.86	0.129	974.33	33.30
	Ours	**100%**	**21.89**	**0.011**	**185.73**	**30.31**
Skateformer [57]	I-FGSM	83.87%	36.66	0.022	521.52	54.50
	MI-FGSM	84.23%	36.83	0.022	523.93	54.83
	MIG	78.14%	57.59	0.037	692.53	78.87
	SMART	65.59%	50.84	0.028	364.26	**13.54**
	Ours	**100%**	**18.67**	**0.011**	**181.89**	28.09
FR-Head [58]	I-FGSM	96.06%	169.34	0.124	1902.39	177.91
	MI-FGSM	96.77%	169.96	0.126	1920.98	179.56
	MIG	92.47%	215.88	0.206	2757.84	243.96
	SMART	81.72%	211.88	0.173	1574.95	55.46
	Ours	**99.28%**	**20.80**	**0.013**	**182.96**	**31.07**
Average	I-FGSM	90.15%	124.69	0.089	1284.58	125.94
	MI-FGSM	90.41%	124.94	0.090	1289.95	126.45
	MIG	84.86%	167.51	0.150	1917.32	187.78
	SMART	72.85%	144.99	0.112	991.50	35.83
	Ours	**99.82%**	**21.07**	**0.012**	**184.47**	**30.41**

Physiological naturalness reflects whether adversarial motions conform to natural movements from the biomechanical perspective of muscle activation. Our method achieves the best performance by a large margin, as shown in Table I. This indicates that even after adversarial attacks, our generated motions retain plausible movements that real-world humans can physically perform. To further evaluate naturalness, as illustrated in Fig. 4, we present the mean power spectral density (PSD) of adversarial samples. The spectral density of our post-attack motions in the high-frequency domain is significantly lower than that of other adversarial motions. Besides, our spectral density closely aligns with the ground truth curve. This suggests that our adversarial motions suffer from few noise-like perturbations because we minimize empirical risk over the data distribution rather than a single sample.

We further compare the effect of explicitly regulating motion dynamics with respect to a given sample in SMART and implicitly regulating motion dynamics through data distribution in our diffusion-based method. Fig. 5 displays the acceleration changes of a sample for comparison. Our method introduces modifications that are not only smaller in magnitude but also smoother and more consistent, whereas SMART does not perturb the motion coherently and consistently. These results imply that minimizing empirical risk over only a given sample leads to deviations from the natural motion distribution and results in the post-attack quality decline.

In terms of physical plausibility, the foot sliding ratio and bone length variation in Table I demonstrate that our generated adversarial motions exhibit superior movement coherency and skeletal consistency compared to other methods. The foot skating ratio indicates the coherence of movements concerning foot contact with the ground. Our method achieves the lowest foot skating ratio, suggesting that post-attack movements are still well synchronized. This is attributed to our method's ability to minimize the risk gap through the integration of data distribution rather than relying on a single sample. Additionally, our adversarial motions exhibit the lowest variation in bone lengths, maintaining consistent skeletal structures across frames. By leveraging access to the data distribution rather than individual samples, our method ensures that modifications preserve cross-frame skeletal integrity.

We also achieve the best performance on the HDM05 dataset, as shown in Table II. Our method successfully fools all systems with an average success rate of 99.82%. This further validates the efficacy of our attack method even when the data quality and amount are not as high as that of the 100STYLE dataset, demonstrating its threatfulness to such systems. Moreover, the distribution of our adversarial motions closely aligns with the ground truth dataset, evidenced by the lowest FID and MMD scores. This validates that our adversarial motions are as natural as pre-attack motions. Besides, physiological naturalness indicates that our generated motions adhere to real-world biomechanical constraints. Since the HDM05 dataset data have been centered to the origin of the coordinates, we do not report the foot skating ratio for this dataset. Additionally, our method achieves superior motion plausibility with the lowest bone length variation. Both the best physiological and physical plausibility demonstrate that our method preserves the quality of adversarial motions.

c) Human Imperceivability: We visualize the trajectories of joints in adversarial motions generated by our method and previous methods in Fig. 6 against the four victim models. In real-world natural movements, joint trajectories are typically smooth and stable. As shown in Fig. 6, our method produces the most stable and smooth trajectories for all joints in the adversarial motions. By minimizing the risk gap over the

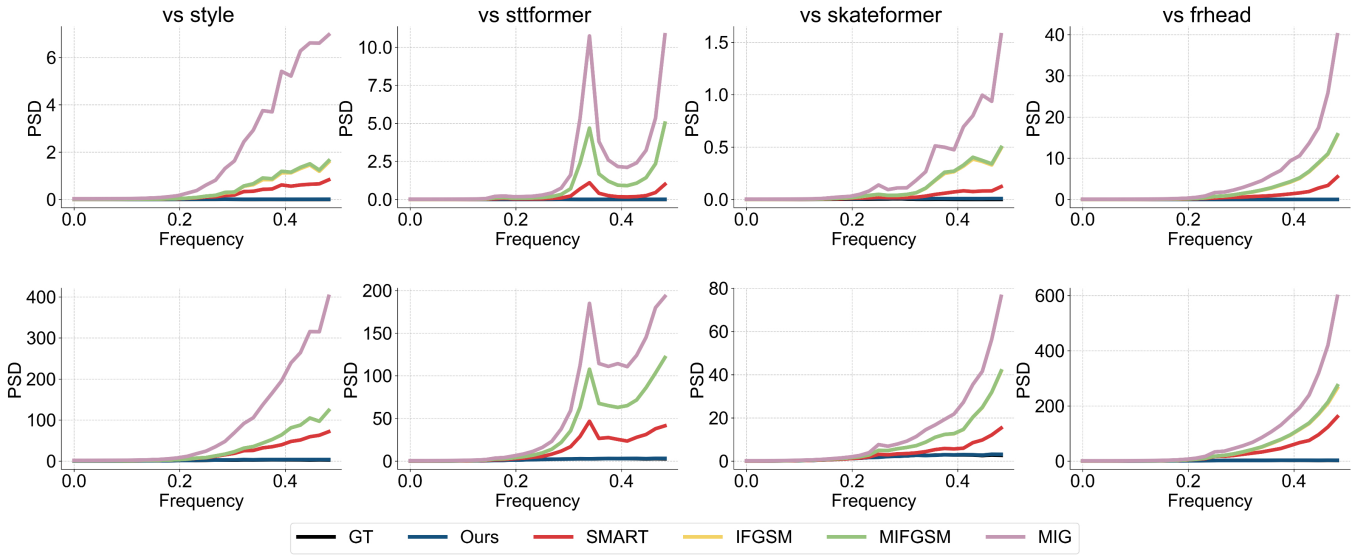


Fig. 4: The mean power spectral density of adversarial samples found on 100STYLE (upper row) and HDM05 (lower row) against four classifiers.

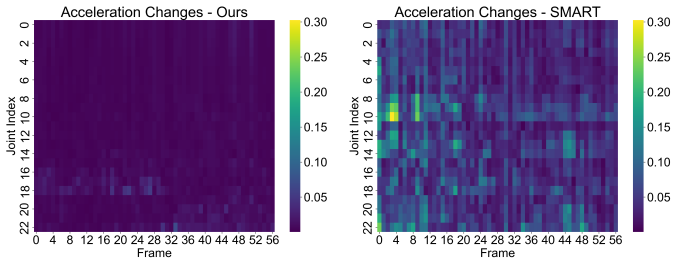


Fig. 5: The visualization of acceleration changes.

data distribution rather than a single motion, our optimization results do not introduce noise-like perturbations, whereas other methods suffer from significant gaps by optimizing over a single input. Consequently, the noise-like perturbations lead to a decline in the post-attack motion quality with observable unstable trajectories and undermines imperceptibility.

Additionally, we recruited volunteers without visual impairments from diverse backgrounds and a balanced gender representation. They participate in questionnaires to evaluate the imperceptibility of our adversarial samples. We provided participants with batches of motions from the same label, which pre-attack motions and post-attack motions generated by our method are mixed together. Participants are asked to select the motions they consider most likely to have potentially been attacked without any time constraints. As illustrated in Fig. 7, our method produces adversarial motions that are the least perceivable by humans.

C. Ablation Study

Given that our method is based on the latent space of a diffusion model, we first perform an ablation study to investigate the influence of the chosen timesteps in constructing the stochastic latent space. Secondly, we examine the impact of different latents used for adversarial attacks to determine which latent is the most suitable for the quality-preserving imperceptible attack task. Finally, we validate the choice of diffusion models by exploring alternative generative models for quality-preserving adversarial attack.

1) Timestep of Stochastic Latent Space: We investigate the impact of different timestep ranges used to map motions from data space to diffusion latent space for adversarial attacks. It has been shown that the semantics encapsulated by the latents have smooth transitions with respect to the diffusion timesteps [66], [67], we conduct experiments using two representative timestep ranges: the earliest timestep range ($t \sim [1, 20]$) and the latest timestep range ($t \sim [980, 1000]$) for comparison.

A trade-off exists between motion naturalness and plausibility, as shown in Table III. This trade-off effect arises from the different emphasis on semantics encoded in the latents. Generally, latents derived from earlier timesteps primarily capture low-level details, while those from later timesteps focus on high-level structural patterns [68]. Utilizing earlier timesteps to construct the stochastic distributional latents is biased toward detailed movements, with a reduced understanding of global coherency and consistency. This bias enhances naturalness by conveying motion dynamics. However, it also exacerbates foot skating and bone length variation due to the lack of global rationality in the latents. Conversely, constructing the distributional latents using later timesteps undermines naturalness, as the generated adversarial motions deviate from the natural distribution, leading to higher FID and MMD scores.

2) Alternative Latents for Attack: We examine the influence of different latents used for our quality-preserving adversarial attacks. Specifically, we compare the posterior mean with the predicted pre-attack data as the latent for adversarial attack. The posterior mean $\boldsymbol{\mu}$ evaluates the required perturbations based on a single timestep of denoising, and allows for step-by-step motion modifications. In contrast, the predicted pre-attack data $\hat{\mathbf{x}}_0$ represents modifications along the entire chain of previous timesteps with approximation. We conduct experiments using the alternative latent to determine whether the posterior mean with a finer-grained, step-by-step evaluation

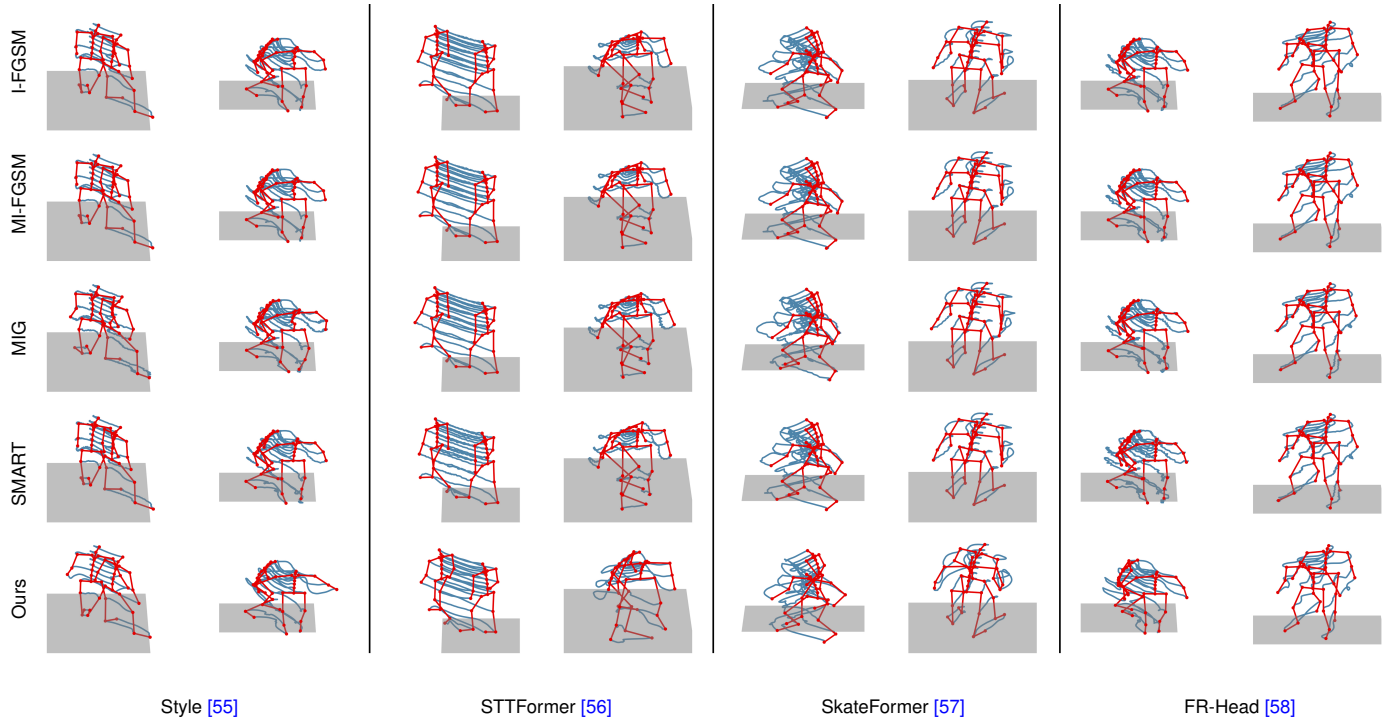


Fig. 6: Visual comparison among the adversarial motions generated by different attack methods against victim models. We visualize the starting and the ending poses in red, the trajectories of all joints in blue, and the ground floor in grey. Our adversarial motions exhibit the most natural and stable trajectories.

TABLE III: The quality of adversarial motions generated via different variants and configurations.

Variants	Configuration	Success Rate ↑	FID ↓	MMD ↓	Physiological Naturalness ↓	Foot Skating ↓	Bone Variation ↓
Timestep	[1, 20]	100%	14.01	0.007	13.28	0.090	3.22
	[980, 1000	100%	26.14	0.017	30.37	0.050	1.62
Latent	$\boldsymbol{\mu}_t$	100%	18.91	0.011	16.05	0.075	2.94
	$\hat{\mathbf{x}}_0$	100%	19.26	0.011	56.70	0.083	3.31
Architecture	Diffusion [64]	100%	18.91	0.011	16.05	0.075	2.94
	VAE [48]	96.06%	198.00	0.056	1058.08	0.083	214.75

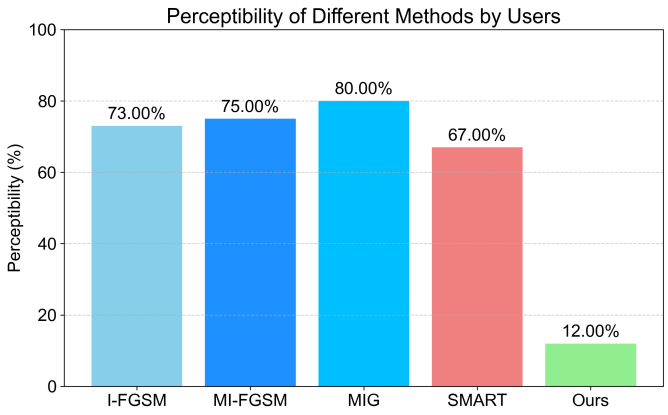


Fig. 7: Perceptibility comparison across different methods.

is more effective to preserve post-attack motion quality than leveraging the predicted pre-attack data over the entire chain.

As shown in Table III, using the predicted pre-attack data as the latent leads to a consistent decline in all performances of adversarial attack. Although both latent choices achieve the same success rate, the quality of adversarial motions deteriorates significantly when switching from the posterior mean to the predicted data. This deterioration stems from the approximation error when calculating the required latents. The predicted $\hat{\mathbf{x}}_0$ is derived from the $\mathbf{x}_{t+1}$ by approximating all previous timesteps $\{i\}_{i=1}^{t}$ collectively, thereby ignoring information from intermediate timesteps. Conversely, the posterior mean μ_t, also derived from the $\mathbf{x}_{t+1}$, considers only the desired changes within a single timestep. This finer-grained modification enhances the quality preservation during adversarial attacks by integrating fewer approximation errors.

3) Alternative Distribution Modeller: Given that our attack method is based on generative models, we explore the use of alternative generative architectures. Similarly, we utilize their latent spaces to modify motions adversarially. Specifically, we conduct experiments using either diffusion models or variational autoencoders (VAEs). We train the VAE with label conditions following the architecture and training configurations outlined in [48]. To maintain consistency with our use

of the posterior mean as latents, we employ the mean of the latent distribution in the VAE. All other configurations remain consistent with those used in our diffusion-based method.

As shown in Table III, the performance of adversarial attacks using a VAE model significantly declines compared to our diffusion-based method. It indicates that effective motion modification requires a comprehensive representation of underlying patterns. In contrast to diffusion models, VAEs utilize a single latent feature rather than a hierarchy of features. Consequently, using a single latent feature in VAEs results in ambiguity regarding motion dynamics and leads to underexpressed movement coherency and consistency.

VII. Conclusion

We propose a novel attack application that imperceptible adversarial motions are achieved without compromising post-attack motion quality. Additionally, we introduce a distribution-based adversarial attack method targeting skeleton-based human action recognition (S-HAR) systems by minimizing the optimization gap inherent in previous approaches. Our method integrates a generative diffusion model, wherein the posterior mean of single timestep denoising is constructed as the proxy to fulfill our attack strategy. To faithfully assess the naturalness of adversarial motions, we develop a new metric aligned with human perception of natural real-world human movements. We evaluate the quality of adversarial motions in terms of threatfulness, motion quality, and imperceptibility, demonstrating that our adversarial motions achieve superior performance across these metrics. The success of our proposed quality-preserving attack application and distribution-based attack method raises significant concerns regarding the robustness of action recognizers, highlighting the necessity for further enhancements in this area.

While achieving natural adversarial motions, our approach presents opportunities for future research. As shown in our experiment, there exists a trade-off in motion quality with respect to the timesteps of the diffusion model. Future work can potentially investigate the influence of trading-off post-attack motion quality on different S-HAR systems. Our proposed physiological naturalness focuses on the physiological perspective of motion naturalness. Future work can potentially integrate the research field of action quality assessment [69], [70] and character animation [71] for more comprehensive motion quality measurements as well as constraints.

References

[1] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *Cyborg and Bionic Systems*, vol. 5, p. 0100, 2024.

[2] Y. Diao, B. Wu, R. Zhang, A. Liu, X. Wei, M. Wang, and H. Wang, "Tasar: Transferable attack on skeletal action recognition," *arXiv preprint arXiv:2409.02483*, 2024.

[3] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 4312–4321, survey Track.

[4] B. Chander, C. John, L. Warrier, and K. Gopalakrishnan, "Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness," *ACM Computing Surveys*, 2024.

[5] H. Wang, F. He, Z. Peng, T. Shao, Y.-L. Yang, K. Zhou, and D. Hogg, "Understanding the robustness of skeleton-based action recognition under adversarial attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 656–14 665.

[6] J. Liu, N. Akhtar, and A. Mian, "Adversarial attack on skeleton-based human action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, pp. 1609–1622, 2020.

[7] N. F. Troje, "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns," *Journal of vision*, vol. 2, no. 5, pp. 2–2, 2002.

[8] S. Shimada and K. Oki, "Modulation of motor area activity during observation of unnatural body movements," *Brain and cognition*, vol. 80, no. 1, pp. 1–6, 2012.

[9] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[10] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[11] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[13] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[14] A. Li, Y. Wang, Y. Guo, and Y. Wang, "Adversarial examples are not real features," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] K. Karunratanakul, K. Preechakul, E. Aksan, T. Beeler, S. Suwajanakorn, and S. Tang, "Optimizing diffusion noise can serve as universal motion priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1334–1345.

[16] L. Kovács, B. M. Bódis, and C. Benedek, "Lidpose: Real-time 3d human pose estimation in sparse lidar point clouds with non-repetitive circular scanning pattern," *Sensors*, vol. 24, no. 11, p. 3427, 2024.

[17] J. Xu, Y. Guo, and Y. Peng, "Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 561–570.

[18] S. Shin, J. Kim, E. Halilaj, and M. J. Black, "Wham: Reconstructing world-grounded humans with accurate 3d motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2070–2080.

[19] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in neural information processing systems*, vol. 4, 1991.

[20] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

[21] G. Claeskens and N. L. Hjort, "Model selection and model averaging," *Cambridge books*, 2008.

[22] Y. Ueyama, "Costs of position, velocity, and force requirements in optimal control induce triphasic muscle activation during reaching movement," *Scientific Reports*, vol. 11, no. 1, p. 16815, 2021.

[23] Y. P. Ivanenko, R. E. Poppele, and F. Lacquaniti, "Five basic muscle activation patterns account for muscle activity during human locomotion," *The Journal of physiology*, vol. 556, no. 1, pp. 267–282, 2004.

[24] I. Mason, S. Starke, and T. Komura, "Real-time style modelling of human locomotion via feature-wise transformations and local motion phases," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, no. 1, may 2022.

[25] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," Universität Bonn, Tech. Rep. CG-2007-2, June 2007.

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[27] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1214–1223.

[28] X. Hu, S. Li, Q. Ying, W. Peng, X. Zhang, and Z. Qian, "Establishing robust generative image steganography via popular stable diffusion," *IEEE Transactions on Information Forensics and Security*, 2024.

[29] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse adversarial perturbations for videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8973–8980.

[30] T. Stolik, I. Lang, and S. Avidan, "Saga: Spectral adversarial geometric attack on 3d meshes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4284–4294.

[31] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[32] Y. Li, W. Jin, H. Xu, and J. Tang, "Deeprobust: A pytorch library for adversarial attacks and defenses," *arXiv preprint arXiv:2005.06149*, 2020.

[33] F. Karim, S. Majumdar, and H. Darabi, "Adversarial attacks on time series," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3309–3320, 2020.

[34] Z. Lu, H. Wang, Z. Chang, G. Yang, and H. P. Shum, "Hard no-box adversarial attack on skeleton-based human action recognition with skeleton-motion-informed gradient," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4597–4606.

[35] Y. Diao, H. Wang, T. Shao, Y. Yang, K. Zhou, D. Hogg, and M. Wang, "Understanding the vulnerability of skeleton-based human activity recognition via black-box attack," *Pattern Recognition*, vol. 153, p. 110564, 2024.

[36] N. Tanaka, H. Kera, and K. Kawamoto, "Adversarial bone length attack on action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2335–2343.

[37] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[38] Z. Chang, G. A. Koulieris, and H. P. Shum, "On the design fundamentals of diffusion models: A survey," *arXiv preprint arXiv:2306.04542*, 2023.

[39] R. Liu, W. Zhou, T. Zhang, K. Chen, J. Zhao, and K.-Y. Lam, "Boosting black-box attack to deep neural networks with conditional diffusion models," *IEEE Transactions on Information Forensics and Security*, 2024.

[40] R. Liu, D. Wang, Y. Ren, Z. Wang, K. Guo, Q. Qin, and X. Liu, "Unstoppable attack: Label-only model inversion via conditional diffusion model," *IEEE Transactions on Information Forensics and Security*, 2024.

[41] C. Hu, Y. Li, Z. Feng, and X. Wu, "Towards transferable attack via adversarial diffusion in face recognition," *IEEE Transactions on Information Forensics and Security*, 2024.

[42] Y. Chen, X. Li, X. Wang, P. Hu, and D. Peng, "Diffilter: Defending against adversarial perturbations with diffusion filter," *IEEE Transactions on Information Forensics and Security*, 2024.

[43] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang, "Human motion generation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[44] S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature reviews neuroscience*, vol. 2, no. 8, pp. 561–567, 2001.

[45] E. Grossman, M. Donnelly, R. Price, D. Pickens, V. Morgan, G. Neighbor, and R. Blake, "Brain areas involved in perception of biological motion," *Journal of cognitive neuroscience*, vol. 12, no. 5, pp. 711–720, 2000.

[46] C. H. Wu and F. De la Torre, "A latent space of stochastic diffusion models for zero-shot image editing and guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7378–7387.

[47] P. Jaini, K. Clark, and R. Geirhos, "Intriguing properties of generative classifiers," in *The Twelfth International Conference on Learning Representations*, 2024.

[48] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 985–10 995.

[49] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 619–10 629.

[50] K. Kim and J. C. Ye, "Noise2score: tweedie's approach to self-supervised image denoising without clean images," *Advances in Neural Information Processing Systems*, vol. 34, pp. 864–874, 2021.

[51] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[52] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2206–2217.

[53] M. Chiquier and C. Vondrick, "Muscles in action," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 091–22 101.

[54] D. Schneider, S. Reiß, M. Kugler, A. Jaus, K. Peng, S. Sutschet, M. S. Sarfraz, S. Matthiesen, and R. Stiefelhagen, "Muscles in time: Learning to understand human motion in-depth by simulating muscle activations," *Advances in Neural Information Processing Systems*, 2025.

[55] L. Zhong, Y. Xie, V. Jampani, D. Sun, and H. Jiang, "Smoodi: Stylized motion diffusion model," in *European Conference on Computer Vision*. Springer, 2025, pp. 405–421.

[56] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," *arXiv preprint arXiv:2201.02849*, 2022.

[57] J. Do and M. Kim, "Skateformer: Skeletal-temporal transformer for human action recognition," in *European Conference on Computer Vision*. Springer, 2025.

[58] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 608–10 617.

[59] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5152–5161.

[60] C. Duan, Z. Zhang, X. Liu, Y. Dang, and J. Yin, "Physics-constrained attack against convolution-based human motion prediction," *Neurocomputing*, vol. 575, p. 127272, 2024.

[61] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[62] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[63] W. Ma, Y. Li, X. Jia, and W. Xu, "Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4630–4639.

[64] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023.

[65] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of riemannian geometry," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 129–24 142, 2023.

[66] A. Sclocchi, A. Favero, and M. Wyart, "A phase transition in diffusion models reveals the hierarchical nature of data," *arXiv preprint arXiv:2402.16991*, 2024.

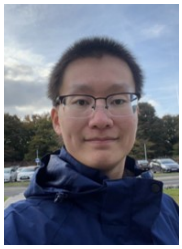
[67] Y. Huang, J. Wang, Y. Shi, B. Tang, X. Qi, and L. Zhang, "Dreamtime: An improved optimization strategy for diffusion-guided 3d generation," in *The Twelfth International Conference on Learning Representations*, 2023.

[68] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," in *The Eleventh International Conference on Learning Representations*, 2023.

[69] K. Zhou, L. Wang, X. Zhang, H. P. H. Shum, F. W. B. Li, J. Li, and X. Liang, "Magr: Manifold-aligned graph regularization for continual action quality assessment," in *Proceedings of the 2024 European Conference on Computer Vision*, ser. ECCV '24. Springer, 2024.

[70] K. Zhou, R. Cai, Y. Ma, Q. Tan, X. Wang, J. Li, H. P. Shum, F. W. Li, S. Jin, and X. Liang, "A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2456–2466, 2023.

[71] H. Wang, E. S. Ho, H. P. Shum, and Z. Zhu, "Spatio-temporal manifold learning for human motions via long-horizon modeling," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 1, pp. 216–227, 2019.



Ziyi Chang is a PhD student in the Department of Computer Science at Durham University. His research focuses on diffusion models with human motions. Specifically, his research involves diffusion models for styled skeleton-based human motion synthesis, skeleton-based human motion analysis, skeleton-based human interaction modelling. He also has interest in 3D surface reconstruction and domain adaptation. He received MSc degree from the University of Edinburgh in 2020 and BSc degree from Renmin University of China in 2019.

Kanglei Zhou received a BSc degree from the College of Computer and Information Engineering at Henan Normal University in 2020. Now, he is pursuing a Ph.D. degree at the School of Computer Science and Engineering, Beihang University. He was also a visiting student with the Department of Computer Science, Durham University, from February to August 2024. His research interests include human motion analysis and augmented reality.

Xiaohui Liang received his Ph.D. degree in computer science and engineering from Beihang University, China. He is currently a Professor, working in the School of Computer Science and Engineering at Beihang University. His main research interests include computer graphics and animation, visualization, and virtual reality.

Hubert P. H. Shum (Senior Member, IEEE) is a Professor of Visual Computing and the Director of Research of the Department of Computer Science at Durham University, specialising in modelling spatio-temporal information with responsible AI. He is also a Co-Founder and the Co-Director of Durham University Space Research Centre. Before this, he was an Associate Professor/Senior Lecturer at Northumbria University and a Postdoctoral Researcher at RIKEN Japan. He received his PhD degree from the University of Edinburgh. He chaired conferences such as Pacific Graphics, BMVC and SCA, and has authored over 180 research publications.