Human Motion Unlearning

Edoardo De Matteis*[1], Matteo Migliarini*[1], Alessio Sampieri[2], Indro Spinelli[1], Fabio Galasso[1]

[1]Sapienza University of Rome
[2]ItalAI

Abstract

We introduce Human Motion Unlearning and motivate it through the concrete task of preventing violent 3D motion synthesis, an important safety requirement given that popular text-to-motion datasets (HumanML3D and Motion-X) contain from 7% to 15% violent sequences spanning both atomic gestures (e.g., a single punch) and highly compositional actions (e.g., loading and swinging a leg to kick). By focusing on violence unlearning, we demonstrate how removing a challenging, multifaceted concept can serve as a proxy for the broader capability of motion "forgetting." To enable systematic evaluation of Human Motion Unlearning, we establish the first motion unlearning benchmark by automatically filtering HumanML3D and Motion-X datasets to create distinct forget sets (violent motions) and retain sets (safe motions). We introduce evaluation metrics tailored to sequential unlearning, measuring both suppression efficacy and the preservation of realism and smooth transitions. We adapt two state-of-the-art, training-free image unlearning methods (UCE and RECE) to leading text-to-motion architectures (MoMask and BAMM), and propose Latent Code Replacement (LCR), a novel, training-free approach that identifies violent codes in a discrete codebook representation and substitutes them with safe alternatives. Our experiments show that unlearning violent motions is indeed feasible and that acting on latent codes strikes the best trade-off between violence suppression and preserving overall motion quality. This work establishes a foundation for advancing safe motion synthesis across diverse applications.

Website — https://www.pinlab.org/hmu

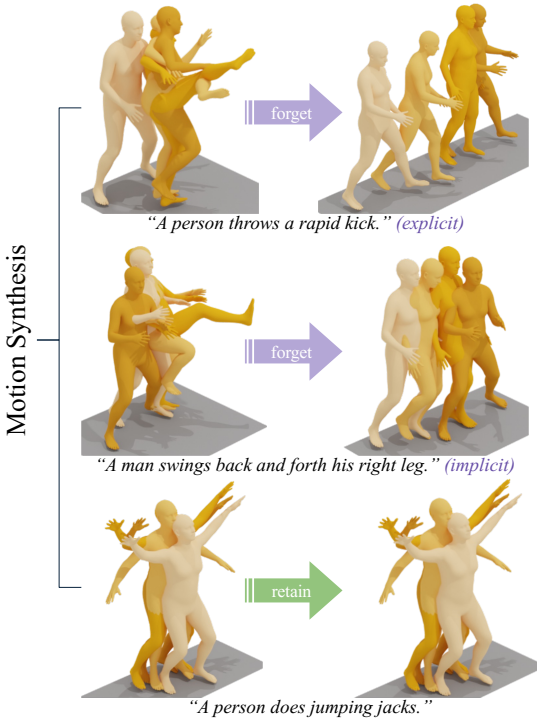


Figure 1: The text-to-motion model takes an input prompt and generates the corresponding motion. With unlearning, when violent content is prompted, the model avoids generating harmful actions, producing a safe and appropriate outcome.

1 Introduction

Generative models have seen dramatic progress across multiple modalities: images (Rombach et al. 2021; Ruiz et al. 2022), videos (Fei et al. 2023; Blattmann et al. 2023), music (Copet et al. 2023), and, more recently, 3D human motions (Zhang et al. 2023; Chen et al. 2023; Sampieri et al. 2024; Jiang et al. 2023). Text-to-motion (T2M) systems now offer unprecedented realism and control, enabling applications in virtual reality (Tirinzoni et al. 2024), animation, and the development of embodied agents and robots trained on human motion data (Kim et al. 2025; Shao et al. 2025; Serifi et al. 2024). However, since popular training corpora such as HumanML3D (Guo et al. 2022) and Motion-X (Lin et al. 2023) encode examples of violent actions (e.g., punching, kicking, stomping), current T2M models can reproduce aggressive behaviors on demand, raising serious safety and ethical concerns. Uncontrolled generation of violent motions risks misuse in simulated environments, flawed human–robot interactions, and unintended bias in downstream controllers that inherit such behaviors (Xu et al. 2023; Merel et al. 2017; Wang et al. 2025).

To address this gap, we introduce the task of Human

*These authors contributed equally.

Motion Unlearning (HMU), with a focus on violence unlearning. Violence serves as an ideal case study as it spans atomic gestures (e.g., a single punch) to highly compositional sequences (e.g., a kick followed by a grapple), requiring fine-grained suppression without degrading non-violent sub-motions. Beyond its role in mitigating real-world risks, such as preventing the synthesis of harmful behaviors in animation and robotics (Merel et al. 2017; Yashuai, Valls, and Dongheui 2023; Wang et al. 2025), violence unlearning also provides a stringent benchmark for motion unlearning more broadly, demonstrating the feasibility of removing specific concepts from a trained generative model.

While we recognize that violent motions are legitimate for specific professionals (e.g., animators), our work focuses specifically on the safety of open models. We address the risks of generating harmful content when these models are distributed to the public, particularly in sensitive fields like robotics. We select violence as our case study because it is both prevalent and conceptually complex in existing datasets, making it an effective benchmark for a safety-oriented unlearning task. Other sensitive motions are too rare for systematic evaluation. Crucially, our approach is not a blunt filtering mechanism: since unlearning operates at the parameter level, it enables developers to release a "safe" public version of a model while retaining a full, unrestricted version for professional use.

We propose a dedicated benchmark for violence-free motion generation based on two recent, large-scale datasets: HumanML3D and Motion-X. From each corpus, we filter out sequences annotated with aggressive movements (punching, kicking, beating, stabbing, etc.) to produce a violence-free subset for evaluation. To capture the sequential nature of motion, our benchmark also includes a violent-only subset and a suite of motion-unlearning metrics that assess both the degree of violence suppression and the preservation of realism and smoothness in transitions between censored and uncensored segments.

Building on state-of-the-art (SotA) image unlearning techniques, we adapt two training-free methods, UCE (Gandikota et al. 2024) and RECE (Gong et al. 2024), to leading T2M architectures (MoMask (Guo et al. 2024) and BAMM (Pinyoanuntapong et al. 2024)). We also include a fine-tuning–based baseline, ESD (Gandikota et al. 2023), for comparison. Drawing inspiration from discrete latent spaces now ubiquitous in motion generation, we introduce Latent Code Replacement (LCR), a novel, training-free approach that identifies violent codes in a VQ-VAE's codebook (van den Oord, Vinyals, and Kavukcuoglu 2017) and substitutes them with safe alternatives, effectively erasing harmful behaviors while maintaining motion fidelity.

Our contributions are threefold:

- **Human Motion Unlearning.** We formulate the novel task of unlearning unsafe motion concepts from pretrained T2M models, contextualized through the challenge of violence prevention in motion synthesis.
- **Violence Unlearning Benchmark.** We curate violence-filtered versions of HumanML3D and Motion-X and define metrics tailored to sequential motion unlearning, establishing a standardized framework to evaluate both suppression efficacy and motion quality.
- **Exploration of Training-Free Methods and LCR.** We adapt UCE and RECE to T2M architectures, propose Latent Code Replacement as a new training-free paradigm leveraging discrete codebooks, and demonstrate that unlearning is indeed feasible; latent-code–based interventions offer the most promising performance, though substantial room for improvement remains.

2 Related Works

Human motion unlearning represents a convergence of machine unlearning and motion synthesis, two active research areas that have remained largely independent until now.

Machine Unlearning. Machine unlearning aims to erase unwanted knowledge from generative models without compromising overall capabilities. This field has gained significant attention in image synthesis (Sai et al. 2024; Xu et al. 2023), where approaches fall into three main categories: data removal methods that retrain models on filtered datasets (Guo et al. 2019; Chien, Pan, and Milenkovic 2022), fine-tuning approaches that adapt selected parameters (Gandikota et al. 2023; Lu et al. 2024; Fan et al. 2024), and training-free interventions that modify model behavior without additional training (Gandikota et al. 2024; Gong et al. 2024).

Data removal quickly becomes impractical for large-scale corpora due to the sheer volume of samples and the prohibitive cost of manual annotation. Among fine-tuning methods, ESD (Gandikota et al. 2023) has been particularly influential. It updates a copy of the pretrained model in a contrastive fashion to penalize the generation of unsafe content. More recently, training-free techniques like UCE (Gandikota et al. 2024) and RECE (Gong et al. 2024) have emerged as SotA solutions. These methods rely on closed-form optimization to map undesirable concepts to predefined targets. Due to their efficiency and strong results, training-free methods now dominate the state of the art in image unlearning. However, despite their success in image generation, these techniques have not yet been extended to sequence-based generative tasks, such as human motion synthesis, where the sequential and structured nature of the data introduces new challenges. In this work, we bridge this gap by adapting training-free unlearning methods to human motion synthesis and by proposing the first benchmark for motion unlearning.

Motion Synthesis. Text-to-motion generation has made significant progress in recent years (Petrovich, Black, and Varol 2022; Tevet et al. 2023; Jiang et al. 2023), fueled by the availability of large-scale motion datasets and advances in deep generative modeling. Modern motion synthesis approaches rely on either continuous (Chen et al. 2023; Sampieri et al. 2024) or discrete latent representations (Guo et al. 2024; Cho et al. 2024; Zhang et al. 2023), with discrete latent spaces emerging as the dominant paradigm in SotA systems. Models like MoMask (Guo et al. 2024) and BAMM (Pinyoanuntapong et al. 2024) use VQ-VAE code-

books (van den Oord, Vinyals, and Kavukcuoglu 2017) to compress motion sequences into discrete tokens, enabling transformer architectures to efficiently capture long-range temporal dependencies and achieve superior scalability and generation quality. Despite the maturity of motion synthesis research, motion unlearning remains entirely unexplored. In this work, we introduce the first dedicated approach to motion unlearning, leveraging the discrete latent structure used in SotA models, and establish a benchmark for evaluating its effectiveness.

Ethics in Generative Models. Recent work has highlighted the importance of safety considerations in generative models across modalities (Dixon et al. 2018; Bansal et al. 2022). Cultural differences in content perception and the challenges of defining "harmful" content universally have been noted in image generation contexts. Our work extends these considerations to the motion domain while acknowledging similar limitations.

3 Human Motion Unlearning

Human Motion Unlearning is the task of selectively removing specific types of motions from trained text-to-motion models while preserving generation quality on acceptable behaviors. Unlike image unlearning, HMU must address temporal dependencies where harmful patterns emerge across sequences rather than in static frames.

We formalize this task as follows. Let $\mathcal{D} = (t_i, m_i)_{i=1}^{N}$ denote a dataset of text-motion pairs, where $t_i \in \mathcal{T}$ are textual prompts and $m_i \in \mathcal{M}$ are corresponding human motions. A T2M model f_θ, parameterized by θ, is trained to map text to motion:

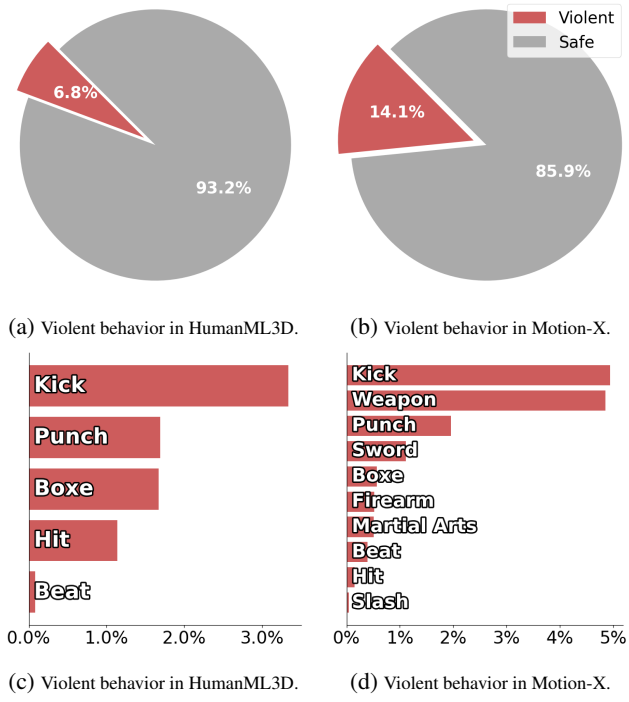
$$f_\theta\left(t \sim \mathcal{T}\right) = m \sim \mathcal{M}. \tag{1}$$

We partition the dataset into a forget set $\mathcal{D}_f$ containing target concepts to be removed, and a retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ containing motions to be preserved. Motion unlearning seeks to reparameterize $\theta \to \theta'$ such that the updated model $f_{\theta'}$ no longer generates samples resembling the distribution of $\mathcal{D}_f$, while maintaining generation quality on $\mathcal{D}_r$.

An additional complication arises from implicit prompting, as shown in Figure 3, a phenomenon also found in images (Yang et al. 2024). Models may generate unwanted motions even when the prompt lacks harmful language. For example, while "*a man throws a punch*" is explicitly violent, but a seemingly harmless alternative like "*a man pulls his arm back and then swings it forward*" may produce the same output. These indirect descriptions effectively function as prompt injections, bypassing safety mechanisms by decomposing a violent act into a sequence of innocuous submotions.

Benchmarking Violence Unlearning

While HMU applies broadly, we focus on violent motions as our instantiation because: **i.** Text-to-motion (T2M) models can generate violent motions with serious social consequences, as motion synthesis datasets like HumanML3D (Guo et al. 2022) and Motion-X (Lin et al. 2023) contain violent behaviors, **ii.** violence offers rich complexity from atomic actions to compositional behaviors, and **iii.** as we will describe shortly, it provides measurable benchmarks for rigorous evaluation. This focus enables us to establish fundamental principles that generalize to other harmful content types.



(a) Violent behavior in HumanML3D. (b) Violent behavior in Motion-X.

(c) Violent behavior in HumanML3D. (d) Violent behavior in Motion-X.

Figure 2: Analysis of violent actions in HumanML3D and Motion-X. (Top row) Pie charts represent the proportion of harmful actions within each dataset. (Bottom row) Bar plots break down the occurrence of individual violent actions.

Datasets. We build our benchmark on two motion datasets: HumanML3D (Guo et al. 2022) is a widely used, large-scale dataset for text-to-motion (T2M) generation, containing approximately 15K motion-capture sequences paired with textual descriptions. Despite its popularity, it includes violent actions such as kicking, punching, and general combat behaviors. Our analysis shows that 7.7% of the motions involve violence (Figure 2a), with kicking (3.4%) and punching (1.7%) being the most frequent (Figure 2c). While the remaining 92.3% of samples are considered safe, the presence of violent behaviors poses risks when these are generated in unintended contexts. Motion-X (Lin et al. 2023) has much larger scale and diversity. It contains around 81K sequences collected from a mix of sources, including videos scraped from the web. Notably, it has a much higher proportion of violent content, 14.9% of its motions are classified as violent (Figure 2b). These include kicking (4.9%), use of weapons (5.0%), and martial arts techniques like Kung Fu and Taekwondo (Figure 2d). This highlights the growing risk of harmful behaviors being embedded as motion datasets scale, particularly when sourced from uncurated, real-world content. Although Motion-X is less acces-

sible and differs in format from HumanML3D, we process it into a compatible format and plan to release it alongside our benchmark.

For both datasets, we define a forget set $\mathcal{D}_f$ containing only violent motions, used to ensure the model forgets harmful behaviors, and a retain set $\mathcal{D}_r$, containing only safe motions, used to verify that the model preserves the quality of the safe content generation. This filtering process is based on a predefined set of violent keywords w, which we curate manually and have already presented in Figures 2c and 2d. A text-motion pair (t, m) is assigned to $\mathcal{D}_f$ if at least one keyword w_i appears in any prompt within $\mathcal{T}$. Since motion is inherently sequential, violent sequences may contain both violent and non-violent sub-motions. Our benchmark assesses whether a model can suppress only violent segments while maintaining smooth transitions and general realism.

To evaluate robustness against implicit prompting, we use GPT-4 to rewrite explicit prompts into subtler, yet semantically equivalent formulations.

Metrics

We align with standard practices in text-to-motion and use FID to assess generation quality, Multimodality Distance (MM-Dist) and R-Precision to evaluate semantic alignment between text and motion, Diversity to measure variation across generations for different prompts, and Multimodality (MM) to measure variation among generations conditioned on the same prompt. When evaluating performance on the forget set $\mathcal{D}_f$, we require a metric that ensures the model avoids generating violent motions. Traditional text-motion alignment metrics fall short here: a model that generates unrelated, random motions would achieve a better score than one that generates correct non-violent submotions while deliberately avoiding violent ones. To address this, we modify MM-Dist by masking out violent words in the text prompts. "A person gives a *kick*" becomes "A person gives a ***". The resulting metric, called MM-Safe, evaluates whether the generated motion remains coherent with the non-violent parts of the prompt. On the retain set $\mathcal{D}_r$, MM-Safe naturally reduces to standard MM-Dist, as there are no violent components to censor.

The following sections detail the state-of-the-art T2M models selected for evaluation and the unlearning methods adapted for our benchmark.

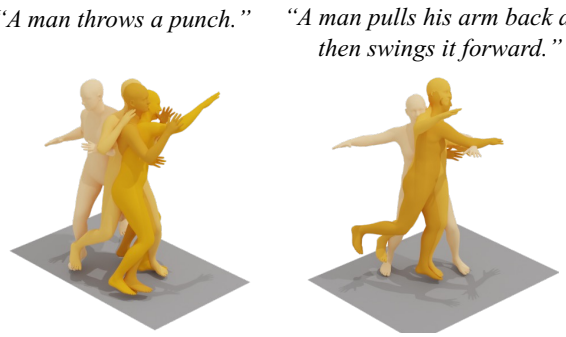


Figure 3: Explicit vs. implicit prompting of violent actions.

4 Text-to-Motion

Text-to-motion synthesis aims to generate 3D motion sequences from natural language descriptions of actions. Recent SotA models rely on discrete latent representations built using VQ-VAEs (van den Oord, Vinyals, and Kavukcuoglu 2017), which enable efficient token-based modeling of motion. Among them, MoMask (Guo et al. 2024) and BAMM (Pinyoanuntapong et al. 2024) currently lead in performance.

Discrete Motion Representation. VQ-VAEs encode an input motion sequence $m \sim \mathcal{M}$ into a continuous latent sequence $Z = E(m) \in \mathbb{R}^{T \times d}$, where T is the motion length and d the embedding dimension. Each vector z_t is quantized to its nearest codebook entry c_{k_t} from a learned codebook $\mathcal{C} = \{c_n\}_{n=1}^{N}$ as:

$$k_t = \arg\min_j \|z_t - c_j\|_2. \tag{2}$$

This results in a discrete latent sequence $Z_q(m) = [k_1, \ldots, k_T]$, which is decoded by $D(\cdot)$ to reconstruct the original motion. Training optimizes three loss terms: a reconstruction loss, a commitment loss, and a codebook update loss (Zhang et al. 2023; Jiang et al. 2023; Guo et al. 2024).

Text-Motion Alignment. Once the motion representation is learned, T2M models align it with textual descriptions to ensure semantic consistency. At inference, given a prompt t, a transformer generates a sequence of latent indices Z_q autoregressively. This discrete sequence is then decoded using the VQ-VAE decoder to produce the final motion sequence.

5 Unlearning Strategies

A growing trend in image machine unlearning is the transition from trainable to training-free methods. While the ideal solution would involve retraining a model from scratch on only safe data, this is typically infeasible due to the high computational cost, time demands, and the need for access to the complete training set. As a result, fine-tuning emerged as a practical alternative, enabling targeted forgetting without full retraining. Recently, however, the state of the art has shifted towards training-free approaches, which achieve superior results with greater efficiency and speed. Most of these methods operate by reparametrizing cross-attention weights; we reimplement these methods in the human motion domain. We also introduce a new perspective: instead of editing attention, we propose to redefine unlearning in the discrete latent space, characteristic of modern text-to-motion architectures.

Attention-Driven Unlearning

ESD. Gandikota et al. (2023) fine-tunes the model's weight matrices W in a contrastive manner, pushing unwanted concepts e_f away from a learned embedding space while pulling them toward a safe, predefined target $\bar{e}_r$.

$$W^{new} \leftarrow W^{old} e_f - \eta(W^{old} \bar{e}_f - W^{old} \bar{e}_r), \tag{3}$$

where η is a guidance scale. Despite the rise of training-free methods, we include ESD and other trainable approaches in our benchmark for completeness.

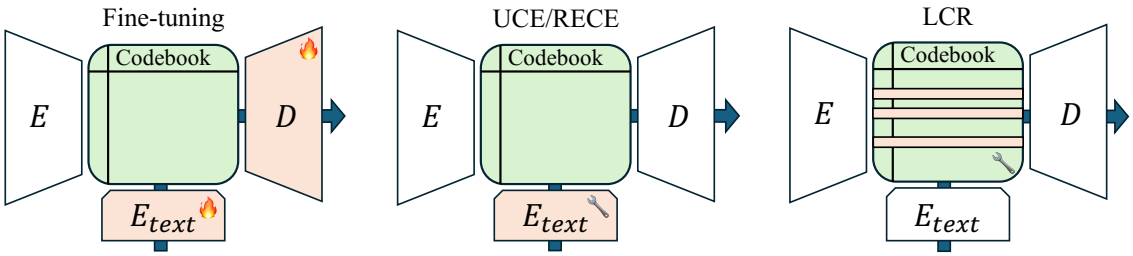


Figure 4: Illustration of motion unlearning approaches: (1) Fine-tuning modifies both the text encoder and motion decoder to remove violent actions, (2) UCE and RECE, as training-free methods, operate solely on the text encoder, (3) Our proposed LCR selectively updates only the affected codebook entries, ensuring targeted unlearning with minimal impact on overall synthesis quality.

UCE. Gandikota et al. (2024) introduces a training-free solution by modifying the cross-attention projection matrix between text and latent representations. Given a set F of concepts to forget, it aligns text embeddings $e_f \in F$ (e.g., "punch someone") to a target embedding $\bar{e}_r$ (e.g., the empty string), which has to be chosen a priori. UCE also retains other safe concepts in R.

$$\begin{aligned}\min_{W^{new}} \sum_{e_f \in F} ||W^{new} e_f - W^{old} \bar{e}_r||_2^2 \\ + \lambda_1 \sum_{e_r \in R} ||W^{new} e_r - W^{old} e_r||_2^2 \\ + \lambda_2 ||W^{new} - W^{old}||_F^2.\end{aligned} \tag{4}$$

Hyperparameters λ_1 and λ_2 control the extent to which retained concepts are preserved. As shown in Gandikota et al. (2024), a closed-form solution exists for Eq. 4.

RECE. Gong et al. (2024) extends UCE to improve robustness. UCE fails to fully forget a concept, as new embeddings e'_f close to the forgotten one e_f still trigger unwanted generations. RECE addresses this by iteratively searching for e'_f near e_f and applying UCE on them. This procedure is repeated for a predefined number of steps, progressively eliminating residual traces of the concept from the latent space.

Latent Code Replacement in Human Motion

Unlike previous approaches that modify attention parameters (Gandikota et al. 2024; Gong et al. 2024; Lu et al. 2024), LCR operates directly on the codebook's discrete latent space for precise representational (see Figure 4). Our method relies on two key assumptions: **i.** motion codes represent disentangled actions, and **ii.** violent motions are identifiable. The first assumption follows from VQ-VAE's theory, where discretization inherently encourages concept disentanglement (Tamkin, Taufeeque, and Goodman 2023). Empirical evidence supports this: removing violent codes preserves generation quality, while injecting them can compromise safe motions (see Appendix C). This targeted latent intervention minimizes disruption to learned parameters while effectively eliminating violent patterns and maintaining high-quality generation for safe motions.

Detecting and Replacing Violent Motion Codes. Given a trained codebook $\mathcal{C}$, let $c \propto \mathcal{D}f$ indicate correlation between code $c \in \mathcal{C}$ and samples in $\mathcal{D}_f$. We want to identify the set of "forget codes" $\mathcal{C}_f = \{c \in \mathcal{C} \mid c \propto \mathcal{D}_f\}$. For each code index k, we compute:

$$N_k(\mathcal{D}) = \sum_{m \in \mathcal{D}} \mathbb{1}\{k \in Z_q(m)\} \tag{5}$$

$$s_k = \frac{N_k(\mathcal{D}_f)}{N_k(\mathcal{D}_r)}, \tag{6}$$

where $N_k(\mathcal{D})$ counts code index k's occurrences in dataset $\mathcal{D}$, and $Z_q(m)$ represents codes activated by motion m. We select the top-K codes with the highest s_k values, which are those that frequently appear in violent motions but rarely in safe ones, to form the violent codebook $\mathcal{C}_f$. Algorithm 1 operates directly on the discrete latent space: each identified violent code is replaced with a randomly sampled safe code $\bar{c} \in \mathcal{C} \setminus \mathcal{C}_f$ plus noise ε to ensure replacement uniqueness. This redirection prevents violent motion generation by substituting violent tokens with safe alternatives during the autoregressive generation process.

6 Results

In this section, we evaluate MoMask and BAMM on HumanML3D and Motion-X using all unlearning strategies from the previous section. Then, we provide qualitative results by comparing generated images from our method to

Algorithm 1: Latent Code Replacement (LCR)

Require: Trained codebook $\mathcal{C}$, forget dataset $\mathcal{D}_f$, retain dataset $\mathcal{D}_r$, number of codes to replace K.
Ensure: Modified codebook with unlearned violent concepts

Compute $N_k(\mathcal{D}_f), N_k(\mathcal{D}_r) \quad \forall k \in \mathcal{C}$
$s_k \leftarrow \frac{N_k(\mathcal{D}_f)}{N_k(\mathcal{D}_r)}$
$\mathcal{C}_f \leftarrow \texttt{Top-}K(s_k)$
$\bar{c} \leftarrow \texttt{sample}(\mathcal{C} \setminus \mathcal{C}_f)$
for each c_f in $\mathcal{C}_f$ **do**
    $c_f \leftarrow \bar{c} + \varepsilon$
end for
return $\mathcal{C}$

		Forget Set			Retain Set			
	Training-Free	FID →	MM-Safe↓	R@1 →	FID ↓	MM-Dist ↓	Diversity →	R@1 ↑
MoMask $\mathcal{D}_r$	✗	$16.358^{\pm.150}$	$4.497^{\pm.018}$	$0.118^{\pm.005}$	$0.075^{\pm.001}$	$2.959^{\pm.002}$	$9.545^{\pm.086}$	$0.512^{\pm.001}$
MoMask	–	$1.164^{\pm.048}$	$5.593^{\pm.073}$	$0.176^{\pm.006}$	$0.041^{\pm.001}$	$2.929^{\pm.002}$	$9.629^{\pm.088}$	$0.520^{\pm.001}$
MoMask *FT*	✗	$2.295^{\pm.065}$	$5.002^{\pm.016}$	$0.150^{\pm.006}$	$0.070^{\pm.001}$	$3.034^{\pm.003}$	$9.680^{\pm.111}$	$0.501^{\pm.002}$
MoMask w/ ESD	✗	$15.039^{\pm.122}$	$6.397^{\pm.033}$	$0.071^{\pm.005}$	$30.679^{\pm.110}$	$7.378^{\pm.011}$	$6.673^{\pm.044}$	$0.165^{\pm.001}$
MoMask w/ UCE	✓	$11.860^{\pm.154}$	$\mathbf{4.626}^{\pm.013}$	$0.135^{\pm.008}$	$0.090^{\pm.001}$	$3.100^{\pm.003}$	$9.733^{\pm.089}$	$0.497^{\pm.001}$
MoMask w/ RECE	✓	$6.952^{\pm.110}$	$4.899^{\pm.016}$	$0.148^{\pm.006}$	$0.144^{\pm.002}$	$3.124^{\pm.004}$	$9.814^{\pm.099}$	$0.493^{\pm.001}$
MoMask w/ LCR	✓	$\mathbf{15.659}^{\pm.128}$	$4.770^{\pm.019}$	$\mathbf{0.124}^{\pm.005}$	$\mathbf{0.050}^{\pm.001}$	$\mathbf{2.986}^{\pm.003}$	$\mathbf{9.523}^{\pm.084}$	$\mathbf{0.508}^{\pm.002}$
BAMM $\mathcal{D}_r$	✗	$12.667^{\pm.256}$	$4.718^{\pm.027}$	$0.112^{\pm.005}$	$0.464^{\pm.005}$	$3.423^{\pm.005}$	$9.762^{\pm.076}$	$0.450^{\pm.002}$
BAMM	–	$0.955^{\pm.055}$	$4.995^{\pm.021}$	$0.180^{\pm.008}$	$0.181^{\pm.003}$	$2.911^{\pm.003}$	$9.731^{\pm.070}$	$0.519^{\pm.001}$
BAMM *FT*	✗	$1.081^{\pm.009}$	$5.000^{\pm.049}$	$0.188^{\pm.011}$	$0.203^{\pm.001}$	$2.948^{\pm.000}$	$9.659^{\pm.004}$	$0.516^{\pm.001}$
BAMM w/ ESD	✗	$0.937^{\pm.030}$	$5.015^{\pm.010}$	$0.194^{\pm.006}$	$0.186^{\pm.001}$	$\mathbf{2.912}^{\pm.001}$	$9.604^{\pm.135}$	$\mathbf{0.521}^{\pm.000}$
BAMM w/ UCE	✓	$38.947^{\pm.773}$	$6.581^{\pm.058}$	$0.088^{\pm.004}$	$0.257^{\pm.003}$	$3.822^{\pm.007}$	$9.480^{\pm.088}$	$0.408^{\pm.002}$
BAMM w/ RECE	✓	$1.428^{\pm.058}$	$4.965^{\pm.030}$	$0.178^{\pm.006}$	$\mathbf{0.170}^{\pm.002}$	$3.137^{\pm.004}$	$9.566^{\pm.059}$	$0.486^{\pm.001}$
BAMM w/ LCR	✓	$\mathbf{7.472}^{\pm.525}$	$\mathbf{4.647}^{\pm.077}$	$\mathbf{0.120}^{\pm.005}$	$0.172^{\pm.001}$	$\mathbf{2.912}^{\pm.001}$	$\mathbf{9.691}^{\pm.166}$	$0.519^{\pm.004}$

Table 1: Results on HumanML3D dataset. We compare various unlearning strategies against the gold-standard $\mathcal{D}_r$, which is trained exclusively on violence-free motions. *FT* denotes the results of fine-tuning the model on the a violence-free dataset. The original MoMask/BAMM models are in grey. → means that the nearer to *Method* $\mathcal{D}_r$ the better.

		Forget Set			Retain Set			
	Training-Free	FID →	MM-Safe↓	R@1 →	FID ↓	MM-Dist ↓	Diversity →	R@1 ↑
MoMask $\mathcal{D}_r$	✗	$9.942^{\pm.488}$	$10.426^{\pm.051}$	$0.172^{\pm.007}$	$11.658^{\pm.124}$	$9.025^{\pm.020}$	$19.869^{\pm.223}$	$0.321^{\pm.002}$
MoMask	-	$6.894^{\pm.338}$	$9.291^{\pm.063}$	$0.322^{\pm.011}$	$3.697^{\pm.062}$	$8.267^{\pm.021}$	$19.343^{\pm.177}$	$0.384^{\pm.003}$
MoMask *FT*	✗	$33.433^{\pm.675}$	$12.838^{\pm.045}$	$\mathbf{0.184}^{\pm.005}$	$4.470^{\pm.046}$	$8.992^{\pm.016}$	$18.485^{\pm.143}$	$0.337^{\pm.002}$
MoMask w/ ESD	✗	$200.89^{\pm1.421}$	$17.977^{\pm.029}$	$0.029^{\pm.004}$	$172.55^{\pm.535}$	$19.001^{\pm.014}$	$6.512^{\pm.076}$	$0.032^{\pm.001}$
MoMask w/ UCE	✓	$53.451^{\pm1.276}$	$14.470^{\pm.070}$	$0.148^{\pm.005}$	$7.252^{\pm.073}$	$10.843^{\pm.026}$	$17.950^{\pm.220}$	$0.275^{\pm.003}$
MoMask w/ RECE	✓	$13.415^{\pm.439}$	$11.205^{\pm.058}$	$0.221^{\pm.008}$	$3.689^{\pm.056}$	$9.142^{\pm.019}$	$19.020^{\pm.182}$	$0.332^{\pm.002}$
MoMask w/ LCR	✓	$\mathbf{7.078}^{\pm.307}$	$\mathbf{9.364}^{\pm.066}$	$0.317^{\pm.009}$	$\mathbf{3.658}^{\pm.060}$	$\mathbf{8.329}^{\pm.017}$	$\mathbf{19.344}^{\pm.184}$	$\mathbf{0.381}^{\pm.003}$

Table 2: Results on Motion-X dataset. The same methods as in Table 1 are used. The original MoMask model is in grey.

those produced by other techniques. We establish an upper bound by training T2M models from scratch on the clean retain set $\mathcal{D}_r$, representing the ideal unlearning performance.

Forget Set Evaluation $\mathcal{D}_f$. Our reference model `Method` $\mathcal{D}_r$ is trained from scratch on the retain set, and expectedly underperforms on violent actions since it never encountered them during training. Unlearned models should closely mimic this upper bound behavior, effectively "forgetting" violent motions while maintaining baseline motion quality.

Retain Set Evaluation $\mathcal{D}_r$. On non-violent motions, we apply standard T2M evaluation metrics (Jiang et al. 2023; Sampieri et al. 2024). Since no violent prompts are involved, models should generate high-quality motions matching the textual descriptions without performance degradation. The MM-Safe reduces to the usual MM-Dist.

HumanML3D. Table 1 (top) evaluates MoMask with LCR unlearning. On the forget set, MoMask w/ LCR achieves the best FID and R@1 while maintaining strong MM-Safe scores, demonstrating effective violence reduction. On the retain set, LCR outperforms all baselines across all metrics, surpassing even MoMask $\mathcal{D}_r$ in FID.

Table 1 (bottom) shows BAMM w/ LCR achieves optimal trade-off between motion quality and violence removal, with Retain FID of 0.172 and MM-Safe of 4.638. Notably, the low forget-set FID values for RECE, ESD, and *FT* indicate failed unlearning of violent motions, whereas LCR's retain metrics closely match MoMask $\mathcal{D}_r$, confirming strong retention of non-forgotten data.

Motion-X. Table 2 presents Motion-X evaluation results. On the forget set, LCR scores 16.3% lower than the second-best method on MM-Safe. On the retain set, LCR closely matches the original MoMask (Guo et al. 2024), preserving better than all the other alternative methods. BAMM unlearning results on Motion-X are provided in Appendix D.

Qualitative Results. Figure 5 shows a qualitative comparison between LCR, UCE, and RECE. In HumanML3D, it is evident that the movements generated by LCR are similar to those of MoMask $\mathcal{D}_r$, where no violence is present. For the actions in Motion-X, it is noticeable that UCE and RECE either hint at violent actions or remain stationary without per-

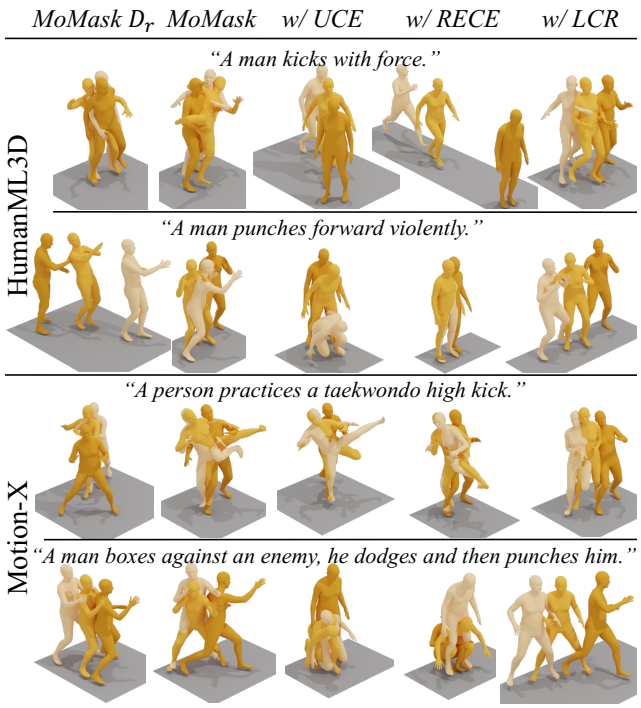


Figure 5: Qualitative comparison across datasets: HumanML3D and Motion-X samples demonstrating unlearning effectiveness. See videos on the project website.

Figure 6: LCR successfully suppresses violent motions even when prompts are rephrased to avoid keywords.

forming any. At the same time, LCR preserves the overall motion without introducing undesirable movements.

7 Discussion

We analyze single-concept unlearning for both violent and non-violent actions, then evaluate robustness against implicit prompting. We focus on LCR due to its superior performance.

Single Concept Unlearning. Table 3 analyzes the impact of removing specific violent actions. We target the most common: *kick* (3.4% in HumanML3D, 4.9% in Motion-X), and the least frequent (*beat* in HumanML3D, *stab* in Motion-X, both less than 0.2%).
Removing *kick* increases FID scores, suggesting reduced toxicity, while maintaining description consistency as MM-Safe remains stable. Also, removing *beat* in HumanML3D is effective (high FID, low MM-Safe), whereas *stab* in Motion-X proves harder to remove (low FID, low MM-Safe). The forget set highlights how the impact varies by action, while the retain set confirms that safe actions remain unaffected, with performance aligning with the baseline.
To showcase broader applicability, we apply the same experiments to non-violent actions, and — in order to avoid any confusion — we provide the complete details, including the updated forget and retain set definitions, in Appendix A.

			Forget Set		Retain Set	
			FID →	MM-Safe↓	FID ↓	MM-Dist ↓
HML3D	$\mathcal{D}_r$	*viol.*	16.358	4.497	0.075	2.959
	LCR	*viol.*	15.659	4.769	0.050	2.986
		kick	17.607	4.827	0.047	2.941
		beat	13.309	4.557	0.201	3.133
MX	$\mathcal{D}_r$	*viol.*	9.942	10.426	11.658	9.025
	LCR	*viol.*	8.331	9.693	3.844	8.704
		kick	36.995	11.222	5.593	9.039
		stab	24.315	7.312	3.624	8.314

Table 3: Single concept unlearning results for MoMask on HumanML3D (*top*) and Motion-X (*bottom*).

Implicit Concept Unlearning. Naive filtering methods based on keyword detection are easily bypassed by rephrasing prompts to avoid banned terms. Since unlearning targets underlying concepts rather than specific words, robust methods must resist such circumvention. As shown in Figure 6, LCR withstands implicit prompting attacks suggesting that concept-level unlearning offers robustness than surface-level filtering. Table 8 in Appendix C compares explicit and implicit evaluations on a set of kicking actions. In the explicit case, our benchmark setup remains unchanged, and LCR outperforms MoMask consistently. In the implicit case, where the violent action is not named in the prompt, MM-Safe becomes equivalent to MM-Dist. Assuming the text embedding still captures the violent intent at the embedding level, a low MM-Dist would imply high alignment with a violent motion, something we want to avoid. Therefore, our goal is to push MM-Dist toward values similar to those of a model trained only on non-violent actions, indicating successful dissociation from violent intent.

8 Conclusions

Text-to-motion models enable key applications but can reproduce violent behaviors from training data. We introduce Human Motion Unlearning to forget selected actions while preserving safe behaviors. We present the first benchmark for motion unlearning with curated datasets and evaluation metrics, and propose Latent Code Replacement, a training-free method that edits motion codes to erase violent content without degrading quality. LCR achieves the best safety-realism trade-off, laying the foundation for safe motion generation and broader unlearning research in generative temporal models.

Ethical Statement

By developing methods to selectively remove violent motions from generative models, our work aims to improve the safety of text-to-motion systems and reduce the computational cost of creating safer models. However, our approach carries potential risks. The capability to manipulate model outputs could be misused to introduce bias, censor legitimate content, or erase representation of specific demographics or cultural practices. Additionally, our definition of "violent" motion reflects particular cultural assumptions that may not generalize across contexts, the same motion could be appropriate in martial arts training but harmful in other settings. While our method demonstrates feasibility, it does not provide complete safety guarantees and may fail to fully suppress targeted concepts or inadvertently degrade overall generation quality. We strongly recommend that text-to-motion models, with or without unlearning, should not be deployed in critical applications without multiple layers of safety controls and careful evaluation of their appropriateness for the specific use case.

Acknowledgements

We thank Luca Franco for the valuable insights and discussions. We acknowledge support from Panasonic, the PNRR MUR project PE0000013-FAIR, and HPC resources provided by CINECA.

References

Bansal, H.; Yin, D.; Monajatipoor, M.; and Chang, K.-W. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? In *EMNLP (Short)*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Chien, E.; Pan, C.; and Milenkovic, O. 2022. Certified Graph Unlearning. In *NeurIPS 2022 Workshop: New Frontiers in Graph Learning*.

Cho, J.; Kim, J.; Kim, J.; Kim, M.; Kang, M.; Hong, S.; Oh, T.-H.; and Yu, Y. 2024. DisCoRD: Discrete Tokens to Continuous Motion via Rectified Flow Decoding. arXiv:2411.19527.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2024. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *The Twelfth International Conference on Learning Representations*.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; and Chua, T.-S. 2023. Dysen-VDM: Empowering Dynamics-Aware Text-to-Video Diffusion with LLMs. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing Concepts from Diffusion Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2426–2436.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified Concept Editing in Diffusion Models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Gong, C.; Chen, K.; Wei, Z.; Chen, J.; and Jiang, Y.-G. 2024. Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models. In *Computer Vision – ECCV 2024: 18th European Conference*.

Guo, C.; Goldstein, T.; Hannun, A. Y.; and van der Maaten, L. 2019. Certified Data Removal from Machine Learning Models. In *International Conference on Machine Learning*.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1910.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating Diverse and Natural 3D Human Motions from Text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795*.

Kim, C. M.; Yi*, B.; Choi, H.; Ma, Y.; Goldberg, K.; and Kanazawa, A. 2025. PyRoki: A Modular Toolkit for Robot Kinematic Optimization. arXiv:2505.03728.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. *Advances in Neural Information Processing Systems*.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024. MACE: Mass Concept Erasure in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Merel, J.; Tassa, Y.; TB, D.; Srinivasan, S.; Lemmon, J.; Wang, Z.; Wayne, G.; and Heess, N. 2017. Learning human behaviors from motion capture by adversarial imitation. arXiv:1707.02201.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*.

Pinyoanuntapong, E.; Saleem, M. U.; Wang, P.; Lee, M.; Das, S.; and Chen, C. 2024. BAMM: Bidirectional Autoregressive Motion Model. In *European Conference on Computer Vision*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sai, S.; Mittal, U.; Chamola, V.; Huang, K.; Spinelli, I.; Scardapane, S.; Tan, Z.; and Hussain, A. 2024. Machine unlearning: an overview of techniques, applications, and future directions. *Cognitive Computation*.

Sampieri, A.; Palma, A.; Spinelli, I.; and Galasso, F. 2024. Length-Aware Motion Synthesis via Latent Diffusion. In *European Conference on Computer Vision*.

Serifi, A.; Grandia, R.; Knoop, E.; Gross, M.; and Bächer, M. 2024. Robot Motion Diffusion Model: Motion Generation for Robotic Characters. In *SIGGRAPH Asia 2024 Conference Papers*, 1–9.

Shao, Y.; Huang, X.; Zhang, B.; Liao, Q.; Gao, Y.; Chi, Y.; Li, Z.; Shao, S.; and Sreenath, K. 2025. Lang-WBC: Language-directed Humanoid Whole-Body Control via End-to-end Learning. arXiv:2504.21738.

Tamkin, A.; Taufeeque, M.; and Goodman, N. D. 2023. Codebook Features: Sparse and Discrete Interpretability for Neural Networks. arXiv:2310.17230.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Tirinzoni, A.; Touati, A.; Farebrother, J.; Guzek, M.; Kanervisto, A.; Xu, Y.; Lazaric, A.; and Pirotta, M. 2024. Zero-shot Whole-Body Humanoid Control via Behavioral Foundation Models. In *NeurIPS 2024 Workshop on Open-World Agents*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*.

Wang, Y.; Yang, M.; Zeng, W.; Zhang, Y.; Xu, X.; Jiang, H.; Ding, Z.; and Lu, Z. 2025. From Experts to a Generalist: Toward General Whole-Body Control for Humanoid Robots. arXiv:2506.12779.

Xu, H.; Zhu, T.; Zhang, L.; Zhou, W.; and Yu, P. S. 2023. Machine Unlearning: A Survey. *ACM Comput. Surv.*

Yang, Y.; Lin, Y.; Liu, H.; Shao, W.; Chen, R.; Shang, H.; Wang, Y.; Qiao, Y.; Zhang, K.; and Luo, P. 2024. Position: towards implicit prompt for text-to-image models. In *Proceedings of the 41st International Conference on Machine Learning*.

Yashuai, Y.; Valls, M. E.; and Dongheui, L. 2023. ImitationNet: Unsupervised Human-to-Robot Motion Retargeting via Shared Latent Space. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, 1–8.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.