

DMP: Directable Motion Retargeting through Motion Paraphrasing

SUNMIN LEE, Seoul National University, Korea

DAVIS REMPE, NVIDIA, USA

YIFENG JIANG, NVIDIA, USA

HAOTIAN ZHANG, NVIDIA, USA

TINGWU WANG, NVIDIA, USA

JUNGDM WON, Seoul National University, Korea

XUE BIN PENG, Simon Fraser University, Canada and NVIDIA, Canada

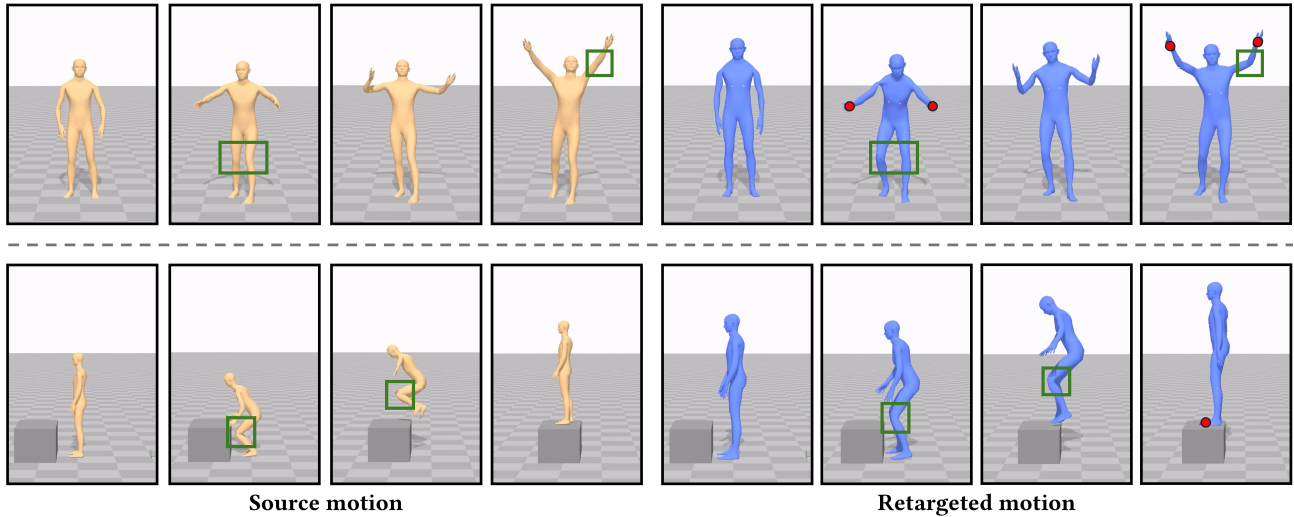


Fig. 1. Directable Motion Paraphrasing (DMP) is a motion retargeting framework that preserves the semantics of a source motion while allowing expressive, user-directed variations. When retargeting an object lifting motion from a short to tall character (top row), the tall character bends its knees and elbows more (green boxes) to account for the hand position constraints (red dots) specified at two object interaction keyframes. For a jumping motion (bottom row), the taller character naturally adjusts its anticipation (green boxes) in response to the foot height constraints (red dots) at the last frame.

Creating diverse and realistic human motions is a fundamental cornerstone of computer animation, with numerous applications in games, movies, and AR/VR. While motion capture is a valuable tool for capturing motions across varied body sizes, obtaining unique motion data for a variety of characters is often prohibitively expensive. Motion retargeting addresses this limitation by adapting existing motions to different character morphologies, however, existing approaches often involve trade-offs between motion realism, user control, and adaptability to artistic needs. In this work, we propose Directable Motion Paraphrasing (DMP), a novel motion retargeting framework based on the concept of *motion paraphrasing*, analogous to text paraphrasing, where the core semantics of a motion are preserved while allowing

expressive, user-directed variations. Our framework constructs a large-scale motion paraphrasing dataset, which captures the diversity of human motion across different body shapes, and trains a diffusion-based generative model that learns both invariances and variations in motion. To enable user control during inference, we introduce a flexible mechanism for specifying spatio-temporal constraints, such as joint positions, rotations, and object interactions, which can be incorporated into the generative process through masked inpainting and loss guidance. We demonstrate the effectiveness of our framework through various examples, showing its ability to produce realistic, diverse, and controllable retargeted motions that meet the artistic demands of animation pipelines. Extensive experiments demonstrate the system's flexibility, motion plausibility, and directability, highlighting its potential as a tool for intuitive and high-quality motion retargeting.

CCS Concepts: • **Computing methodologies** → **Animation; Motion processing; Neural networks.**

Additional Key Words and Phrases: Character Animation, Motion Retargeting, Conditional Motion Generation

ACM Reference Format:

Sunmin Lee, Davis Rempe, Yifeng Jiang, Haotian Zhang, Tingwu Wang, Jungdam Won, and Xue Bin Peng. . DMP: Directable Motion Retargeting through Motion Paraphrasing. *ACM Trans. Graph.* 1, 1 (January), 16 pages. <https://doi.org/10.1145/3829359>

Authors' Contact Information: Sunmin Lee, sunmin.lee@imo.snu.ac.kr, Seoul National University, Korea; Davis Rempe, drempe@nvidia.com, NVIDIA, USA; Yifeng Jiang, yifengj@nvidia.com, NVIDIA, USA; Haotian Zhang, haotianz@nvidia.com, NVIDIA, USA; Tingwu Wang, tingwu@nvidia.com, NVIDIA, USA; Jungdam Won, jungdam@imo.snu.ac.kr, Seoul National University, Korea; Xue Bin Peng, xbpeng@sfu.ac.kr, Simon Fraser University, Canada and NVIDIA, Canada.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© Copyright held by the owner/author(s).
ACM 1557-7368/11-ART
<https://doi.org/10.1145/3829359>

1 Introduction

Creating diverse and realistic human motions is a fundamental cornerstone in computer animation, with numerous applications in games, movies, and AR/VR. A key aspect of this involves accommodating diverse human morphologies, which is essential for creating authentic and compelling characters, and enhancing narrative depth. While motion capture is a valuable tool for capturing motions across varied body sizes, obtaining unique motion data for a variety of characters is often prohibitively expensive. *Motion retargeting*, the process of transferring motions from one character to another, is a standard solution for adapting existing motion capture data to new character morphologies. Motion retargeting is also vital for practical applications of motion synthesis models, which often operate on a canonical character morphology.

The fundamental challenge of motion retargeting is to preserve the salient characteristics of the source motion while generating natural movements appropriate for the target character. This is inherently an ambiguous problem. The best solution depends on what the user considers to be the essential characteristics of the source motion, and how they want these characteristics adapted to the target character’s morphology and behavior. Traditional optimization-based methods resolve this ambiguity through precise user control in the form of constraints, such as joint position and rotation constraints. While this approach provides users with maximum flexibility, it is challenging to maintain motion quality, as the optimization lacks priors for human motion. As a result, achieving realistic results often requires extensive manual intervention, such as specifying dense keyframes or meticulous tuning of constraints. This process is not only labor-intensive, but also demands considerable animation expertise. Alternatively, recent neural retargeting models learn how to naturally adapt motions to different body proportions by leveraging large motion datasets. By training on motions performed by characters with different morphologies, they can generate plausible retargeted motions automatically, requiring minimal manual user intervention. However, these learned methods lack the controllability needed for users to explore different retargeting possibilities to best suit their artistic vision, which is crucial in animation pipelines.

A complete motion retargeting solution should output natural motions for the target character morphology, while preserving the core semantics of a source motion and allowing users to direct the results according to their artistic needs. Therefore, in this work we frame the process of retargeting as *motion paraphrasing*, analogous to text paraphrasing, which takes the core idea of a sentence and expresses it in different vocabularies or sentence structures. Similarly, motion paraphrasing takes a source motion and *reenacts* the motion with a different target character, maintaining the underlying meaning of the source motion while allowing expressive variations that users can direct. For example, motion paraphrasing of the box stepping motion in Fig 1 (bottom) should maintain the stepping height of the character while the exact arm movements may vary.

In this work, we present Directable Motion Paraphrasing (DMP), a motion retargeting framework based on *motion paraphrasing*, which offers a range of possible retargeting solutions, intuitive control over the output motions, and realistic retargeted motions that align with the morphology of a given target character. First, we construct a

large-scale *motion paraphrasing dataset* consisting of motion capture recordings from diverse subjects performing similar motion sequences. Then, we train a generative diffusion-based motion retargeting model to generate plausible variations of a given source motion on different target skeletons. For finer-grained control over the retargeted motions, our model enables users to specify flexible kinematic constraints at inference time. These user constraints are seamlessly incorporated into the diffusion framework through a combination of masked inpainting and loss guidance. We demonstrate the flexibility of the system on a suite of spatio-temporal constraints, such as joint positions, rotations, contact, distance constraints, and object collision avoidance. Furthermore, we conduct an extensive quantitative analysis of crucial design choices to assess their impact on motion quality and directability in our system.

The central contributions of this work are as follows:

- We consider the retargeting problem through the lens of *motion paraphrasing* and propose a practical algorithm for constructing a motion paraphrasing dataset.
- We propose the first directable diffusion-based motion retargeting model.
- We demonstrate the effectiveness and practicality of our method through qualitative and quantitative evaluations.

2 Related Work

We begin by exploring the extensive history of motion retargeting research, with an emphasis on the plausibility of retargeted motions and intuitive user control. We then discuss recent advancements in generative motion models and conditional motion generation, highlighting their relevance for directable motion retargeting, and how our work pushes the boundaries within this research context.

2.1 Motion Retargeting

Motion retargeting has been an active area of research in computer animation since the seminal work of [Gleicher 1998], which formulated motion retargeting as a space-time nonlinear constrained optimization problem. Initial efforts explored techniques for enhancing computational efficiency through methods such as hierarchical motion editing [Lee and Shin 1999], inverse rate control [Choi and Ko 1999], and per-frame Kalman filtering [Tak and Ko 2005]. Subsequent studies have further explored methods for improving the physical plausibility of retargeted motions [Al Borno et al. 2018; Popović and Witkin 1999; Tak and Ko 2005], and modeling advanced kinematic features to effectively capture and preserve intricate spatial relationships [Al-Asqhar et al. 2013; Ho et al. 2010; Liu et al. 2018]. To achieve plausible results, these optimization-based methods often demand extensive manual tuning to define effective optimization objectives for each specific behavior, posing significant challenges for most users. Although Shin et al. [2001] proposed a solution that identifies key characteristics of a given source motion and applies simple inverse kinematics to preserve those features in the retargeted motion, this approach lacks the generality needed to handle diverse behaviors and relies heavily on hyperparameter tuning. In contrast, our retargeting method learns to preserve the semantically salient characteristics of motions from diverse behaviors, without relying on manually designed heuristics.

Recent advances in deep learning have generated significant interest in neural motion retargeting approaches, leveraging motion priors learned from real human motion datasets rather than relying on manually intensive optimization objectives. A fundamental challenge in training effective retargeting models is the scarcity of datasets with paired motions between different characters. Researchers have explored various unsupervised learning methods to address this challenge. Several methods have leveraged indirect supervision through cycle-consistency [Aberman et al. 2020; Hu et al. 2023; Lim et al. 2019; Tripathi et al. 2025; Villegas et al. 2018; Zhang et al. 2024a] and adversarial learning [Hu et al. 2023; Villegas et al. 2018; Zhang et al. 2024b]. These methods often adopt kinematic objectives from optimization-based motion retargeting methods [Aberman et al. 2020; Lim et al. 2019; Villegas et al. 2018; Zhang et al. 2023a] and physics-inspired objectives to incorporate more direct supervision into training, circumventing the challenges arising from the lack of paired datasets [Tripathi et al. 2025]. Several works have focused on improving semantics preservation by explicitly modeling geometric relationships between body limbs and preventing interpenetration [Cheynel et al. 2025; Villegas et al. 2021; Yang et al. 2025; Ye et al. 2024; Zhang et al. 2024a, 2023a], and by further leveraging vision-language models [Zhang et al. 2024b]. As an alternative to unsupervised learning methods, Lee et al. [2023] proposed a supervised learning framework that utilizes synthetic motion pairs generated by an optimization-based retargeting algorithm. As a result, the quality and diversity of this approach are inherently limited by that of the underlying optimization algorithm. While neural approaches can generate plausible retargeted motions automatically with minimal user intervention, they are limited in that users cannot effectively control the outcome of the retargeting. In this work, we propose the first controllable, generative motion retargeting framework. By training a diffusion-based model on a large-scale *motion paraphrasing* dataset that captures diverse retargeting solutions, and leveraging effective inpainting and guidance techniques, our system outputs a variety of plausibly retargeted motions that users can easily direct to their artistic preferences, while preserving the original semantics.

2.2 Generative and Conditional Motion Modeling

Human motions are highly diverse and multimodal. Advances in generative motion modeling have provided powerful tools for capturing this complex distribution including Variational Autoencoders (VAEs) [Ling et al. 2020; Petrovich et al. 2021; Rempe et al. 2021], Generative Adversarial Networks (GANs) [Barsoum et al. 2018; Li et al. 2022], normalizing flows [Henter et al. 2020], and vector quantized models (VQ-VAEs) [Pinyoanunpong et al. 2024; Zhang et al. 2023b]. In recent years, diffusion models [Ho et al. 2020] have proven their exceptional ability to capture complicated distributions from massive datasets and generate high quality samples from various domains including images [Nichol and Dhariwal 2021] and videos [Ho et al. 2022]. Moreover, the iterative nature of denoising makes it amenable to controllability through methods like classifier-free guidance [Ho and Salimans 2022], classifier guidance [Dhariwal and Nichol 2021], gradient-based guidance from various loss functions [Chung et al. 2023], and direct inpainting [Lugmayr et al. 2022]. The success of diffusion has carried into the human motion

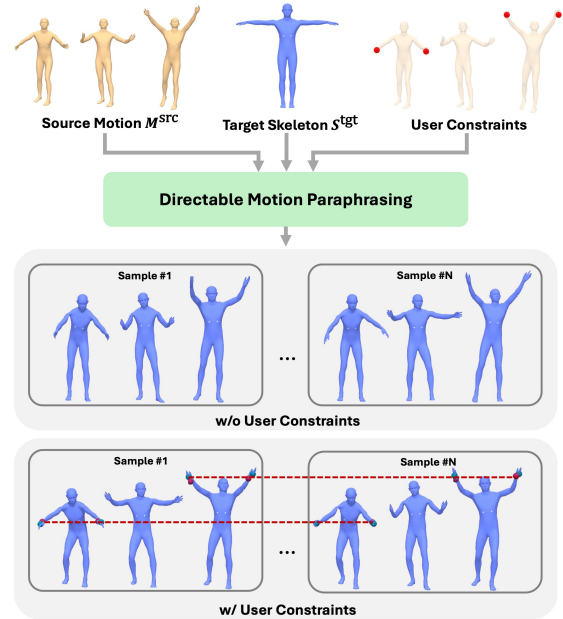


Fig. 2. The user workflow for our directable motion paraphrasing (DMP) model. Starting with a source motion and a target character, DMP generates diverse yet plausible motion retargeting results, illustrated in the top gray box. Users can refine these motions by adding spatial constraints, such as specifying joint positions at particular frames. These user-provided constraints (red dots) further enhance the motion retargeting outcomes, as shown in the bottom gray box. DMP typically achieves satisfactory results within just a few iterations.

domain [Dabral et al. 2023a; Shafir et al. 2023; Tevet et al. 2023; Zhang et al. 2022] where it has enabled generating high-quality motion sequences while offering precise control. Diffusion models have been effectively utilized for various conditional motion generation tasks, such as text-to-motion [Tevet et al. 2023], audio-to-motion [Alexanderson et al. 2023], multimodal gesture synthesis [Ao et al. 2023], interactive real-time character control [Chen et al. 2024; Shi et al. 2024], trajectory and joint conditioned motion synthesis [Karunratanakul et al. 2023; Rempe et al. 2023; Xie et al. 2023], motion inbetweening [Cohan et al. 2024; Studer et al. 2024], motion style transfer [Hu et al. 2024; Raab et al. 2024; Song et al. 2024; Zhong et al. 2025], scene-conditioned motion [Huang et al. 2023; Yi et al. 2024], and text-based motion editing [Athanasios et al. 2024].

We introduce the first diffusion-based retargeting model, leveraging the strength of diffusion in modeling non-intuitive and complex mappings within large-scale datasets, specifically the mapping between semantically consistent motions performed by various actors. Furthermore, the flexibility to control outputs through inpainting and guidance enables us to develop intuitive interfaces for a user-controllable motion retargeting framework.

3 Motion Paraphrasing

In this work, we aim to develop a directable motion retargeting model that automatically preserves the semantically relevant characteristics of a source motion, while also offering users with flexible

and intuitive control and refinement over the resulting motions. To this end, we define the concept of *motion paraphrasing*: retargeting that maintains high-level semantics while accommodating natural variations in low-level details. This concept is inspired by text paraphrasing, where multiple sentence variations can convey the same underlying meaning. Just as writers can adjust individual words and sentence structure while paraphrasing until they are satisfied with the details, we address directable motion retargeting by allowing users to control retargeted results within the distribution of plausible variations. Figure 2 illustrates the user workflow for a motion paraphrasing model. Given a source motion, a target character, and optional constraints, the motion paraphrasing model generates diverse yet plausible motion retargeting solutions, preserving high-level semantics close to those of the source motion. Users can then select a solution from the results or further refine it by modifying existing constraints or specifying additional constraints. Variations in the generated retargeting solutions are analogous to different sentences conveying the same semantics, while adding constraints resembles adding extra words to further specify the meaning in text paraphrasing.

To this end, we propose a data-driven motion paraphrasing framework. Specifically, we first construct a motion dataset comprising various individuals performing the same actions (Sec 4). We then train a diffusion-based motion paraphrasing model to learn the rich distribution of plausible retargeting variations from the motion paraphrasing dataset (Sec 5). After training, users can specify additional spatio-temporal constraints at test-time, which are incorporated into the diffusion framework through masked inpainting and loss guidance (Sec 6). Note that our work focuses on motion retargeting across characters with a fixed skeletal topology, where the body proportions may vary. This scope is intentionally different from prior learning-based retargeting approaches, such as SAN [Aberman et al. 2020], SAME [Lee et al. 2023], and PAN [Hu et al. 2023], which aim to generalize across diverse skeletal structures. Instead, we focus on learning user-directable motion retargeting, allowing natural motion variations while satisfying given constraints.

4 Motion Paraphrasing Dataset

To capture a wide variety of possible motion retargeting solutions for training our model, we constructed a large-scale *motion paraphrasing dataset* comprised of motion capture recordings from diverse subjects performing similar behaviors. When acquiring the dataset, motion capture actors were directed to mimic or enact a particular behavior given visual or text prompts, and each actor performed the behavior multiple times to exhibit natural variations in the motion. We group the recorded motions of the actors according to their respective prompts, producing a dataset containing groups of motion clips from different actors that correspond to the same underlying behaviors. Random pairs of motion clips can then be sampled from each group to produce examples of how actors with different body shapes perform a similar behavior. However, these motion pairs are often temporally misaligned, since each actor may have different motor capabilities and movement styles. The motion clips may also have different speeds and include varying lengths of idle motions

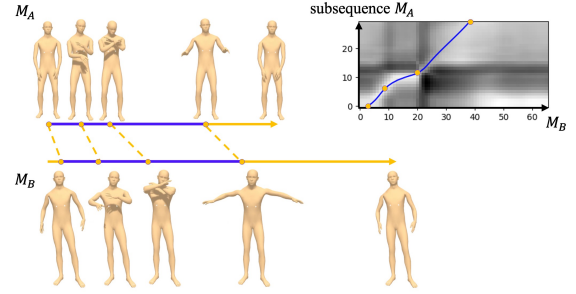


Fig. 3. An example of dynamic time warping to generate temporally aligned motion paraphrasing data pairs. M_A and M_B are motion capture data of similar behaviors performed by different subjects. Given a subsequence of M_A , highlighted in blue, the optimal subsequence within M_B , also in blue, is found through subsequence dynamic time warping. The top right shows the similarity score matrix between subsequence M_A and M_B , where the x-axis represents the time axis of M_B , and the y-axis represents the time axis of a subsequence of M_A . The darker the area is, the higher the cost, and the blue line denotes the optimal match minimizing the total cost.

at the beginning or end of the clip, leading to additional misalignments. As introducing such misalignments is generally undesirable within the retargeting workflow, we temporally filter and align the motion pairs through subsequence Dynamic Time Warping (DTW) [Müller 2015]. At a high level, the algorithm finds the optimal match, after re-timing, between the frames of two sequences based on a user-defined similarity metric. We heuristically define the similarity cost between the i -th pose M_A^i of motion M_A , and j -th pose M_B^j of a second motion M_B , by

$$\begin{aligned} C(M_A^i, M_B^j) = & w_R \|R_A^i \ominus R_B^j\|^2 + w_\omega \|\omega_A^i - \omega_B^j\|^2 \\ & + w_c \|f_A^i - f_B^j\|^2 + w_h \|h_A^i - h_B^j\|^2 \\ & + w_d \|d_A^i - d_B^j\|^2, \end{aligned} \quad (1)$$

where R denotes the joints rotations, ω represents the angular velocities, h refers to the average foot heights, f are the binary foot contact labels, and d denotes the 2D root displacement on the ground plane. The operator $\ominus$ measures the geodesic distance between two rotations. Foot contact serves as an important semantic indicator of a pose, revealing key aspects of a motion, such as the phase of a gait cycle or step sequences of a dancing motion. When climbing stairs or ladders, both the contact label and the height of the feet are important factors for pose comparison. Additionally, the root horizontal displacement direction is compared to better differentiate poses moving in different directions. The displacement direction of each frame is computed by averaging a window of 0.5 seconds. Weights $w_R = 1$, $w_\omega = 2$, $w_c = 0.3$, $w_h = 5$, $w_d = 0.5$ are used to balance the various metrics. If the best-alignment cost across the trajectory is higher than a specified threshold after DTW, this means a good match cannot be found and we discard the motion pair. We segment motion clips into short subsequences and apply subsequence DTW, as large discrepancies between actors are naturally more prevalent when capturing longer motion sequences, making good matches more difficult to find. Specifically, given two motions from the same behavior group, one motion is randomly segmented into a short subsequence, and the optimal *subsequence* within the

other motion that best matches the short segment of the first motion is determined through subsequence DTW. An example is illustrated in Figure 3. Note that, unlike tasks where low-level kinematic metrics are used to discover semantic similarities among unstructured motion datasets, our dataset construction process applies DTW to group of motions where the actors are instructed to perform the same behaviors. Therefore, high-level semantic equivalence is guaranteed by our capture protocol, which greatly simplifies the task for DTW to be responsible for low-level temporal alignment.

5 Motion Paraphrasing Model

To capture the rich diversity of retargeting variations in a motion paraphrasing dataset, we propose a novel diffusion-based motion paraphrasing model. Our model for directable motion paraphrasing (DMP) retargets a given source motion and source skeleton to a motion suitable for a target skeleton, while preserving the underlying semantics of the source motion. Additionally, diverse output motions can be generated by starting with different initial noise. To incorporate directability into the model, random masked inpainting [Cohan et al. 2024] is applied during training, helping the model to learn to respect motion constraints specified at inference time.

5.1 Motion Representation

A motion M is represented as a sequence of frame features, and a feature at frame i consists of the root position $p^{\text{root}} \in \mathbb{R}^3$, root heading direction $d^{\text{root}} = (\cos \phi, \sin \phi) \in \mathbb{R}^2$, where ϕ is the planar rotation angle, joint rotations $\{R^1, \dots, R^J\} \in \mathbb{R}^{J \times 6}$ encoded in the 6D representation [Zhou et al. 2019], joint positions $\{p^1, \dots, p^J\} \in \mathbb{R}^{(J-1) \times 3}$, joint velocities $\{v^1, \dots, v^J\} \in \mathbb{R}^{J \times 3}$, and foot contact labels $f \in \{0, 1\}^4$, where J denotes the number of joints, with the superscript 1 reserved for the root. The root position and heading direction are expressed relative to the planar root coordinate system at the first frame of the motion. The origin of this coordinate system corresponds to the root position projected onto the ground, while the rotation is determined by the heading direction, the world up-vector, and their normalized cross-product. Joint positions are expressed with respect to their current planar root coordinate system, excluding the planar root position since it remains constant within this coordinate system. Foot contact labels f are boolean values that indicate whether feet and toes are in contact with the ground or objects. In our dataset, foot contact labels are computed automatically by checking whether the joint velocity is below a manually specified threshold. Since the motion dataset does not include environment information, proximity to the object cannot be directly measured, and the contact labels are solely determined by the joint velocities.

5.2 Diffusion Model and Training

We choose to use a generative diffusion model for motion paraphrasing due to its ability to capture complex distributions from large datasets while being controllable through inpainting and test-time guidance. Our model is trained following the DDPM framework [Ho et al. 2020]. Given a motion from the data distribution $M_0 \sim q(M)$, a forward diffusion process progressively applies Gaussian noise at

each diffusion step according to the equation

$$q(M_t|M_{t-1}) = \mathcal{N}(M_t; \sqrt{1 - \beta_t}M_{t-1}, \beta_t \mathbf{I}), \quad (2)$$

where the variance β_t follows a schedule designed to transform the data distribution $q(M)$ to an isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as the diffusion steps progress $t \rightarrow \infty$. The denoising process is defined as the reverse process, which iteratively denoises M_t starting from noise $M_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and progressing to the data distribution $M_0 \sim q(M)$. Since the true conditional distribution of the reverse process is unknown, a denoiser is used to approximate it:

$$p_\theta(M_{t-1}|M_t, C) = \mathcal{N}(M_{t-1}; \mu_\theta(M_t, t, C), \Sigma_t \mathbf{I}), \quad (3)$$

where θ is a set of learnable parameters, $\mu_\theta(M_t, t, C)$ is the predicted mean motion of the previous step computed from the noisy motion M_t at the current step, Σ_t is the variance following a predefined schedule for the reverse process, and C denotes a set of additional conditions provided as input to the denoiser. Rather than directly predicting μ_θ , the denoiser π_θ modeled by deep neural networks first estimates the clean motion $\hat{M}_0 = \pi_\theta(M_t, t, C)$. Then μ_θ is computed by combining M_t and the estimation $\hat{M}_0$, following Nichol and Dhariwal [2021]:

$$\mu_\theta(M_t, t, C) = \frac{\sqrt{\alpha'_t} \beta_t}{1 - \alpha'_t} \hat{M}_0 + \frac{\sqrt{\alpha'_t}(1 - \alpha'_{t-1})}{1 - \alpha'_t} M_t, \quad (4)$$

where auxiliary variables are defined as $\alpha_t = 1 - \beta_t$ and $\alpha'_t = \prod_{s=0}^t \alpha_s$. To generate a motion that aligns with a source motion and a given target skeleton, the denoiser network incorporates the source motion M^{src} , the source skeleton S^{src} , and the target skeleton S^{tgt} into the context C . The skeleton's morphology $S \in \mathbb{R}^{J-1}$ is represented by the length of each link.

5.2.1 Conditioning on Constraints. To enable user directability of the retargeted motions at sampling time, the model is trained with randomly sampled constraints as input. During training, constraints are created by randomly sampling a subset of motion features from the ground-truth target motion, such as root position, heading orientation, joint positions, and joint rotations. Constrained features are indicated by a binary mask m which has the same dimension as a motion M . Then, feature constraints from the ground-truth target motion M_0^{tgt} are directly inpainted into the motion currently being denoised M_t^{tgt} as

$$\tilde{M}_t^{\text{tgt}} = m \odot M_0^{\text{tgt}} + (1 - m) \odot M_t^{\text{tgt}}, \quad (5)$$

where $\odot$ refers to element-wise multiplication. The masked motion $\tilde{M}_t^{\text{tgt}}$ is then provided as input to the denoiser, along with the mask m to indicate which features are constrained.

A key design decision when randomly generating constraints during training is the structure of the constraint sampling process. Naïve random sampling of features will lack the structure that is exhibited by user-specified constraints at test time. For example, randomly sampling features of a 6D joint rotation representation independently may impose constraints on only one element, whereas users are more likely to constrain the entire rotation in their actual workflow. Therefore, we designed a structured constraint sampling scheme that better emulates the actual user workflow. We designed a set of commonly used constraint types and randomly sampled

combinations of constraints during training. Constraint types include:

- Fullbody-Keyframes: positions or rotations of all body joints at a specific keyframe.
- Partial-Keyframes: position or rotation of a specific joint at a specific keyframe.
- Dense-Root-Trajectory: root trajectory of all time frames. The heading orientation of root trajectory is also constrained by probability of 0.5.
- Sparse-Root-Keyframes: root position and orientation at a specific keyframe.

For constraints on a specific keyframe (*Fullbody-Keyframes*, *Partial-Keyframes*, *Sparse-Root-Keyframes*), 1-5 random frames per sequence are chosen as keyframes. For constraints on specific joints (*Partial-Keyframes*), joints are sampled from a binomial distribution ($p=0.2$). With a probability of 0.1, we do not impose any constraints, which enables the model to retarget plausible motion in both conditional and unconditional settings.

5.2.2 Architecture and Training. A schematic illustration of our model architecture is shown in Figure 4. A transformer encoder embeds the source motion M^{src} and the source skeleton S^{src} , then a transformer decoder takes the embedded source along with the latest motion being denoised $\tilde{M}_t^{\text{tgt}}$, the target skeleton S^{tgt} , diffusion timestep t , and a motion mask m to predict a clean retargeted motion $\hat{M}_0 = \pi_\theta(S^{\text{src}}, M^{\text{src}}, S^{\text{tgt}}, \tilde{M}_t^{\text{tgt}}, t, m)$. The transformer decoder alternates between self-attention and cross-attention to provide the conditioning signal on the source motion embeddings. During training, a batch of retargeting motion pairs $\{(M^{\text{src}}, M^{\text{tgt}})\}^B$, where B is the batch size, is sampled from the motion paraphrasing dataset. A noise level corresponding to a uniformly sampled diffusion step t is then applied to the retargeted motion through the forward process, resulting in $\tilde{M}_t^{\text{tgt}}$. This noisy motion is then modified as $\tilde{M}_t^{\text{tgt}}$ through Equation (5) with a randomly sampled motion mask m . The denoiser is supervised using the loss function below, which combines a reconstruction loss with a constraint loss,

$$L = w_{\text{rec}} \mathbb{E}_B \left[\|\hat{M}_0 - \tilde{M}_0\|^2 \right] + w_{\text{con}} \mathbb{E}_B \left[\|m \odot (M_0^{\text{tgt}} - \hat{M}_0)\|^2 \right]. \quad (6)$$

where w_{rec} and w_{con} are the weights for the reconstruction and constraint losses, respectively. The constraint loss encourages the model to prioritize constrained features during denoising.

6 Directable Motion Paraphrasing

At test time, random noise is sampled from a standard Gaussian distribution $M_t^{\text{tgt}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and then iteratively denoised given a source motion M^{src} and skeleton S^{src} , target skeleton S^{tgt} , and a motion constraint mask m . To accelerate the generation process, we adopt DDIM sampling [Song et al. 2020]. When no constraints are provided (i.e., m is all zeros), the model generates a distribution of retargeted motions on the target skeleton that are semantically identical to the source, while exhibiting natural variations. Beyond exploring diverse retargeting variations, users can further direct the model by specifying spatio-temporal constraints, such as joint

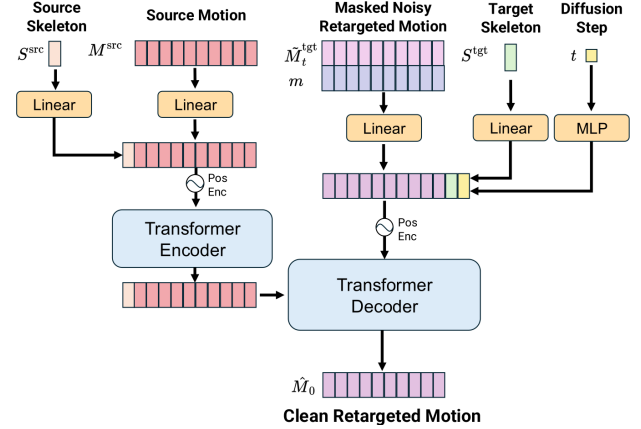


Fig. 4. Our paraphrasing model is implemented using a diffusion model. In each denoising step, the model takes in the source skeleton S^{src} and source motion M^{src} along with the masked noisy motion $\tilde{M}_t^{\text{tgt}}$ and mask m , target skeleton S^{tgt} , and current diffusion timestep t . A transformer decoder predicts the clean retargeted motion $\hat{M}_0$ conditioned on the embedded source motion through cross attention.

positions, rotations, inter-body-part distance, and collision avoidance. These constraints are integrated into denoising in two ways: masked inpainting and loss guidance. The interleaving of inpainting and/or guidance with the denoising steps during sampling preserves the naturalness of motion, which would otherwise be difficult to achieve with traditional optimization-based techniques.

6.1 Masked Inpainting

Users can apply constraints to the retargeting process by directly specifying desired values for various motion features. Similar to training time, the constraint values are directly inpainted into the current noisy motion $\tilde{M}_t^{\text{tgt}}$ at each denoising timestep t of the reverse process before being given to the denoiser,

$$\tilde{M}_t^{\text{tgt}} = m \odot M^{\text{usr}} + (1 - m) \odot \tilde{M}_t^{\text{tgt}} \quad (7)$$

where M^{usr} is a feature vector that includes user-specified values.

6.2 Loss Guidance

In addition to masked inpainting, diffusion guidance is used to enable more accurate enforcement of constraints. Guidance also supports more flexible constraints, which cannot be easily specified through inpainting, such as the spatial relationship between body parts or objects, or collision avoidance. Let $l(M)$ be an analytical loss function that quantifies the difference between the user-provided constraints and the constraints obtained from a (clean) retargeted motion M . To reduce constraint violation, a predicted motion $\hat{M}_0 = \pi_\theta(S^{\text{src}}, M^{\text{src}}, S^{\text{tgt}}, \tilde{M}_t^{\text{tgt}}, t, m)$ is updated by the gradient of the objective function at each denoising step to produce a corrected motion $\tilde{M}_0$,

$$\tilde{M}_0 = \hat{M}_0 - w \nabla_{\tilde{M}_t^{\text{tgt}}} l(\hat{M}_0), \quad (8)$$

where w is the step size for the update. Note, the gradient is computed with respect to the noisy input motion $\tilde{M}_t^{\text{tgt}}$ rather than the

predicted clean motion $\hat{M}_0$ [Ho et al. 2022; Karunratanakul et al. 2023; Rempe et al. 2023]. Compared to computing the gradient with respect to $\hat{M}_0$, which affects only the features directly used for computing the loss l , computing the gradient with respect to M_t allows a sparse constraint to have dense influence on all frames and features by back-propagating gradient through the denoiser model, enabling a more coherent change to the overall motion. We design guidance objectives to constrain joint position, rotation, relative joint position, self-penetration avoidance, object collision avoidance, and object contact.

Below, we provide examples of guidance objective functions that can be incorporated into our system. Note that our method is not limited to these specific guidance objectives, and custom objective functions can be specified according to user requirements.

Global Joint Position. The global positions of joints often convey important semantic information, especially for object and scene interaction behaviors. Let $\hat{R}^j, \hat{p}^j = \text{FK}(\hat{M}_0)$ be the global joint rotation and position of joint j , where FK denotes a forward-kinematics operation. To encourage the resulting motion to reach a target position $\bar{p}^j$ with joint j , the guidance objective can be specified simply with the position error $l_p^j = \|\hat{p}^j - \bar{p}^j\|^2$.

Joint Rotation. Similarly, the guidance objective for reaching a target joint rotation $\bar{R}^j$ can be specified by, $l_R^j = \|\hat{R}^j \ominus \bar{R}^j\|^2$, where $\ominus$ computes the geodesic distance between two rotations. This can be useful when it is desired to copy subtle high-frequency details from source to target motion.

Relative Joint Position. The relative spatial relationship of two joints often conveys important semantic cues that may not be captured through independent joint positions alone. For example, in a drinking motion, a character’s hand should move toward their mouth. To preserve such relative spatial relationship, we use the following objective:

$$l_{rel}^{i,j} = \|\hat{p}^{i,j} - \bar{p}^{i,j}\|^2 + \|\hat{p}^{j,i} - \bar{p}^{j,i}\|^2, \quad (9)$$

where $\hat{p}^{i,j} = (\hat{R}^j)^{-1}(\hat{p}^i - \hat{p}^j)$ represents the relative position of i -th joint with respect to j -th joint. $\hat{R}^j$ and $\hat{p}^j$ denote the global rotation and global position, computed using the same forward-kinematics described above. $\bar{p}^{i,j}$ is the desired relative position, which can be computed from the source motion.

Self-Penetration Avoidance. When the generated motions are applied to characters with meshes, self-penetration may occur, which can undermine the plausibility of the motion. Guidance can also be applied to prevent these self penetration artifacts. First, we approximate a character’s mesh using spheres, and self-penetration is minimized by penalizing penetration depths among the spheres. The penetration depth between i -th sphere and j -th sphere can be computed by $l_{sp}^{i,j} = (\|c^i - c^j\| - (r_i + r_j))^2$, where c^i and c^j are the centers of sphere i and j , respectively, and r_i and r_j are their respective radii. The penetration depths of adjacent spheres that initially overlap in the T-pose, are excluded from the computation.

Object Collision Avoidance. When a character is moving through a scene, it should avoid not only self-penetration but also unnecessary collisions with external objects. We approximate a character’s

occupancy mesh using spheres. If the i -th sphere collides with k -th object, the collision is penalized by $l_{col}^{i,k} = \min(\text{SDF}^k - r_i, 0)^2$, where SDF^k represents the signed distance field of the k -th object, and r_i is the radius of the i -th sphere.

Object Contact. In contrast to avoiding collisions with certain objects, there are also scenarios where contact with specific objects is necessary. For example, hands should touch a box in order to grab it. To encourage the generated motion to correctly contact an object, while also allowing for flexibility in determining the contact positions, a contact objective can be defined by $l_{ct}^{i,k} = c^{i,k}(\text{SDF}^k - r_i)^2$, where $c^{i,k}$ is a binary variable indicating whether the i -th sphere should make contact with the k -th object.

7 Experiments

In this section, we demonstrate the capabilities of our proposed directable motion retargeting framework. In Sec 7.1, we first provide details of our motion paraphrasing dataset and model implementation. Sec 7.2 demonstrates the key capabilities of our system, with examples of retargeting diverse motions while following various user-specified constraints. Sec 7.3 and Sec 7.4 provide quantitative evaluations of the model’s performance compared to baselines and ablations in the constrained and unconstrained settings. Finally, as our dataset is commercially licensed and cannot be released, we also evaluate our model with a dataset constructed from publicly available data (Sec 7.9) to facilitate reproducibility and enable follow-up study.

7.1 Implementation Details

Dataset. Our motion paraphrasing dataset is constructed from a large motion capture dataset containing motion from 170 participants. The participants include approximately equal numbers of male and female, with heights ranging from 145cm to 199cm, roughly following a normal distribution with mean 174.3cm and std 10.9cm (Fig. 11). The dataset includes common motion categories such as locomotion, everyday activities and gestures, and combat with a variety of styles. The raw motion captures vary from 1 to 180 seconds in length. The dataset consists of thousands of distinct behaviors, each performed by multiple actors and for multiple repetitions. These distinct behaviors are split into a non-overlapping train (90%) and test set (10%), such that the test set will not contain any behaviors used during training. Motion pairs are constructed for training using the subsequence dynamic time warping process described in Sec 4, resulting in relatively short motion pairs ranging from 1 to 4 seconds in length. The final motion paraphrasing training set contains 67M motion pairs. To better handle object interactions at test time (e.g., jumping on boxes with variable heights), we augment the training dataset by randomly scaling both the source and target character skeletons by the same scale along with their corresponding root translations. The scale factor is uniformly sampled between $[0.8, 1.2]$.

Model and Training. We implement our framework with PyTorch [Paszke et al. 2017] and PyTorch-Lightning [Falcon and The PyTorch Lightning team 2019]. The source motion transformer encoder consists of 8 encoder layers with 8 attention heads and latent dimension

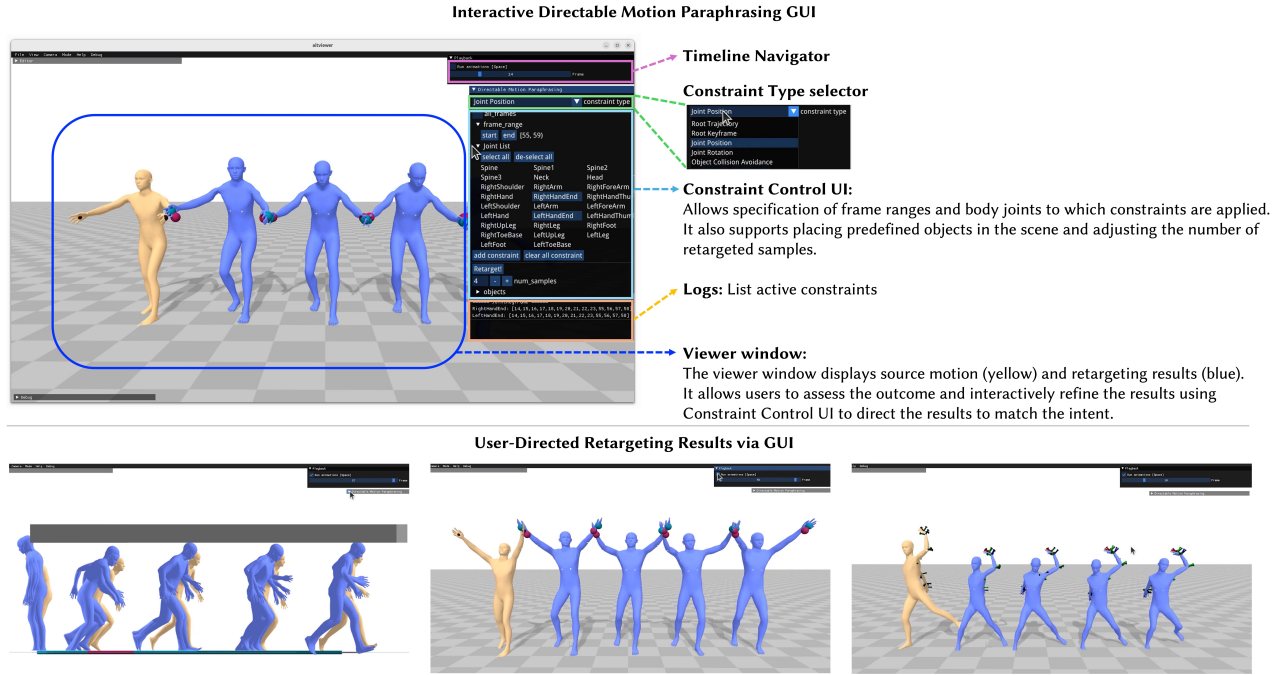


Fig. 5. (Top) The interactive interface developed for our directable retargeting framework, allowing artists to easily specify the type, location, and duration of user constraints with just a few clicks. (Bottom) Examples of retargeting results achieved through our interactive GUI under different user-specified constraints. (Bottom Left) Object collision avoidance constraints are applied to duck under an obstacle. (Bottom Middle) Position constraints (red spheres) are applied to the hands to ensure the character reaches towards a target location. (Bottom Right) Joint rotation constraints are applied to the arms, and the model infers the motion of the remaining joints.

1024, while the transformer decoder is a stack of 16 layers with 8 attention heads and latent dimension 1024. The model has a total of 221M learnable parameters, and was trained with batch size 2048 and learning rate $1e-5$ using 4 NVIDIA A100 GPUs for approximately 6 days.

7.2 Interactive Directable Retargeting

To demonstrate how animators can utilize our directable retargeting framework, we developed an interactive program prototype shown in Fig 5. We used *aitviewer* library for visualization [Kaufmann et al. 2022]. Given a source motion and a target skeleton, users can visualize undirected (i.e., unconstrained) motion samples generated by our diffusion-based paraphrasing model. Then, through an intuitive graphical user interface, users can iteratively add constraints of various types (e.g., joint positions), and regenerate motions to meet specific requirements. Finally, the user can select one or more retargeted motions for final use. The supplementary video includes three demonstrations of this retargeting workflow.

Fig. 6 presents qualitative examples of unconstrained generation from our paraphrasing model for dancing and kicking motions. Given a source motion, our DMP model effectively captures the distribution of natural variations of the same behavior, including subtle arm movements and differences in foot placement, while preserving the core semantics of the motion, such as the side kick. This comprehensive coverage of motion variations is crucial in the

directed setting for diffusion guidance and inpainting: it enables generating in-distribution motion samples that satisfy all user-specified constraints. The supplementary video provides additional animated examples of undirected generation, which demonstrates our model’s ability to preserve the timing of movements between the source motion and retargeted outputs. This timing consistency, a direct result of our careful curation of the paraphrasing dataset, supports common retargeting workflows where maintaining the timings of source motions is oftentimes preferred.

Fig. 7 illustrates examples of DMP generations that are guided by various types of user constraints applied through masked inpainting and loss guidance. Compared to undirected generation from the same model, applying constraints such as object avoidance, wrist joint positions, and arm rotations further enhances the physical plausibility and overall quality of the motion, while aligning more closely with user intent. The supplementary video includes additional animated examples, showcasing the use of other types of constraints, such as maintaining distances between body parts or following a specified root trajectory, across activities like running and lifting a box.

It is worth noting that our dataset is built using standard motion capture, which captures skeletal motion but does not include information on skinning shapes. All visualizations in this paper use a nominal skinning mesh scaled to match the provided bone lengths, which does not fully represent the actual body shape of the actor.

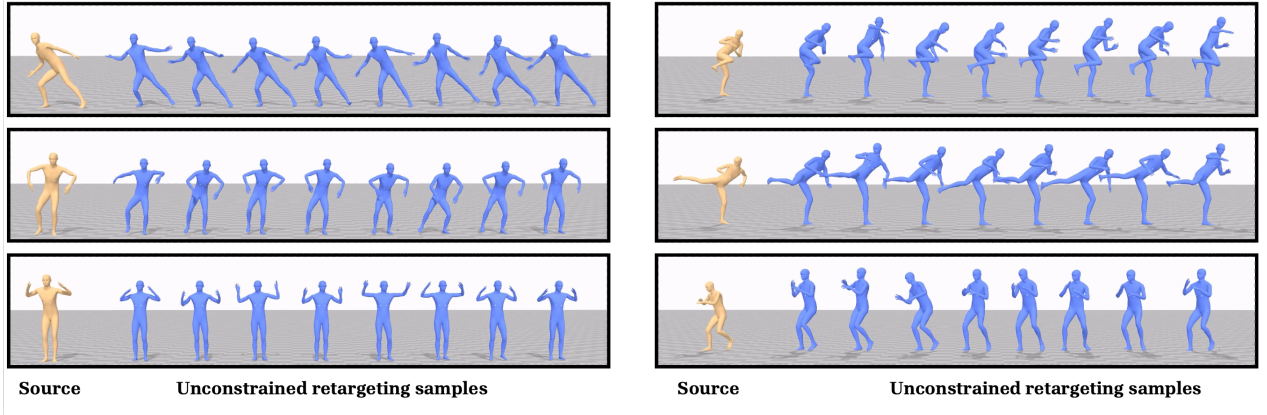


Fig. 6. In the unconstrained retargeting setting, our diffusion-based paraphrasing model generates natural variations of the target skeleton’s behavior while preserving the key semantics of the source motion. Given a source motion and target skeleton as input, the model produces samples, such as a locking dance (left) and a spinning back kick (right), demonstrating its ability to maintain the essence of the original motion.

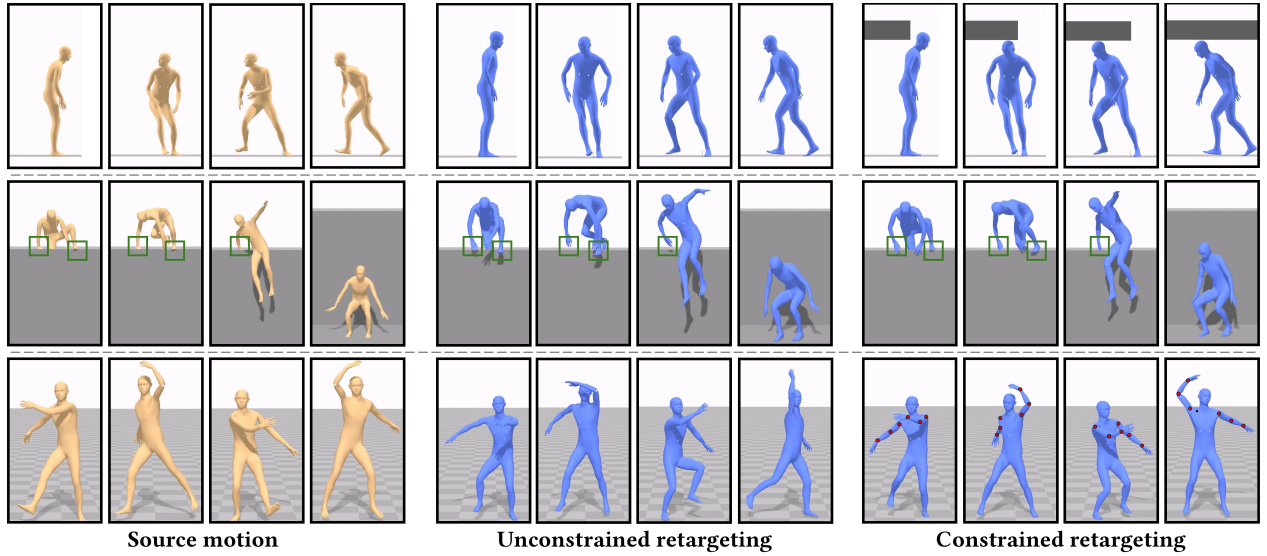


Fig. 7. Besides unconstrained retargeting, our framework enables constrained (directed) retargeting, where the user can specify spatial-temporal constraints at inference time. The example in the first row visualizes a character turning back and running to the left. With added ceiling and object avoidance constraints, the character crouches more to run below the ceiling. The second example shows a character jumping off a wall. With enforced contact constraints added to the hand and foot joints, the character better preserves the contacts (green boxes) from the source motion. The last example depicts a character performing a hip-hop dance. With joint rotation constraints (red dots) specified on the arm joints at every frame, the character dances in a more similar rhythm and magnitude with the source motion.

7.3 Quantitative Results: Constrained Retargeting

To evaluate the directability of our retargeting model, we measure the accuracy of constraint-following in several settings on a test set of 1000 motion pairs containing held out behaviors. For each source motion in the test set, we randomly sample two constraints among root position, root orientation, joint positions, and joint rotations keyframes and sample up to 5 keyframes. Constraint values are determined by the input source motions. Target skeletons are randomly sampled with a height difference of at least 15cm with respect to the source character to eliminate overly simple test cases. The first set

of metrics used to evaluate performance measure the violation with respect to the given constraints. Position errors are reported in cm in the global coordinate frame and rotations are in degrees. Note that joint position constraints may become infeasible when the target character is shorter than the source, which contributes to the joint position error. We also report metrics to evaluate the quality of the generated motions. We assess the distributional similarity between the generated and ground-truth motions using Frechet Inception Distance (*FID*). In line with prior works [Dabral et al. 2023b; Li et al. 2021; Zhang et al. 2025], we compute FID with well-established

Table 1. Constraint violation and motion quality comparison for constrained retargeting. Rotation copy and optimization-based baselines hit constraints accurately, but can produce unnatural motions without additional priors. The best values are shown in **bold**.

Method	Constraint Violation				Motion Quality			
	Root Pos	Heading Orient	Joint Pos	Joint Rot	FID	Foot Skate	Jerk	Penetration
RotCopy-RootScale	6.959	0.000	14.239	0.000	1.387	0.338	0.054	0.021
Opt	0.956	0.007	2.918	0.000	22.174	0.404	0.180	0.024
Optcont	0.930	0.033	3.422	0.000	17.072	0.417	0.078	0.035
DMP (Ours)	2.884	2.304	8.574	3.304	4.290	0.295	0.085	0.016

Table 2. Ablations of our method to evaluate important components of our system, including masked inpainting and guidance at inference time, and different types of masking strategies during training.

Method	Constraint Violation				Motion Quality			
	Root Pos	Heading Orient	Joint Pos	Joint Rot	FID	Foot Skate	Jerk	Penetration
Inference w/o Mask	3.788	2.694	10.797	6.527	4.547	0.299	0.084	0.015
Inference w/o Guide	3.801	2.378	10.516	3.296	4.695	0.292	0.085	0.015
Train w/o Mask	4.130	3.015	11.351	6.808	4.342	0.273	0.072	0.018
Train w/ Random Mask	4.157	2.737	10.653	3.745	6.447	0.240	0.046	0.018
Train w/o ConLoss	2.688	2.328	8.825	4.740	4.832	0.260	0.063	0.019
DMP (Ours)	2.884	2.304	8.574	3.304	4.290	0.295	0.085	0.016

kinetic features that captures dynamic aspects of motions [Onuma et al. 2008], using the *fairmotion* implementation [Gopinath and Won 2020]. *Foot Skate* (m/s) is the mean foot velocity when the foot should be in contact with the ground (based on height from the floor), *Jerk* (km/s³) captures motion smoothness by measuring the change in acceleration, and *Penetration* (cm) is the sum of the distance of all joints below the floor when penetration is detected. Lower is better for all metrics. The best values are highlighted in **bold** in all tables. As generative models yield diverse predictions for the same input, it is common practice to use the minimum metric value from a set of multiple samples [Li et al. 2023; Xu et al. 2024; Yuan and Kitani 2020]. We generate 20 samples for each test cases, and report the minimum. The guidance weights in Eq. (8) are set to $w = 250$. Our motion representation includes both joint rotations and joint positions. This enables the specification of constraints on both joint rotations and positions through feature-wise masking. To visualize and evaluate the motions generated by our model, we use joint positions derived from the predicted joint rotations via forward kinematics to ensure bone length consistency. We find that the model’s joint position and rotation predictions are highly consistent. Without additional loss or guidance on the joint position consistency at test time, joint position predictions and the joint positions derived from FK has an average difference of 0.007m. Similarly, the motion representation also includes joint velocity features, where the joint velocity feature and the joint velocity computed via finite differences of the joint positions are consistent, with an average error of 0.06m/s.

7.3.1 Baseline Comparisons. Table 1 compares our method to multiple retargeting baselines, which are representative of standard heuristics-based methods used in practice. *RotCopy-RootScale* first

copies the joint rotations of a source motion to the target skeleton, then scales the root translation of the source motion by the ratio of root height between the target skeleton and source skeleton in a default standing pose. This baseline trivially achieves 0 error for rotation constraints, but it is unaware of position constraints resulting in much higher position errors. While the reported motion quality constraints are on par with our approach, the naïve translation scaling can result in egregious errors as shown in the supplementary video. The *Opt* baseline applies a gradient-based optimization algorithm using Pytorch-minimize [Feinman 2021] on top of *RotCopy-RootScale*, using the same objective functions as those used for guidance. Optimization allows for high accuracy in satisfying constraints, but has no prior on natural human motion, and therefore leads to large discontinuities and unrealistic motion as indicated by high FID, foot skate and jerk. *OptCont* addresses large discontinuities and jerk by incorporating additional objectives that penalize abrupt pose changes across consecutive frames. However, since *OptCont* does not leverage any motion data, it can produce unnatural behaviors, such as overly smoothed motions, which leads to a high FID. Also note that the results of *OptCont* are subject to a trade-off between constraint accuracy and smoothness, depending on the smoothness regularization weight. The optimal weight can be scenario-dependent and often requires manual tuning, whereas DMP (Ours) performs well across diverse scenarios. Figure 8 visually compares DMP (Ours) with baseline approaches to highlight the difference. DMP generates plausible retargeted motions that satisfy user constraints, whereas optimization-based baselines (*Opt*, *OptCont*) yield large discontinuities or unnatural smoothing, and simple scaling (*RotCopy-RootScale*) fails to satisfy positional constraints.

7.3.2 Ablation Study. To evaluate the impact of our design decisions, we compare our full DMP method (*Ours*) to several ablations,

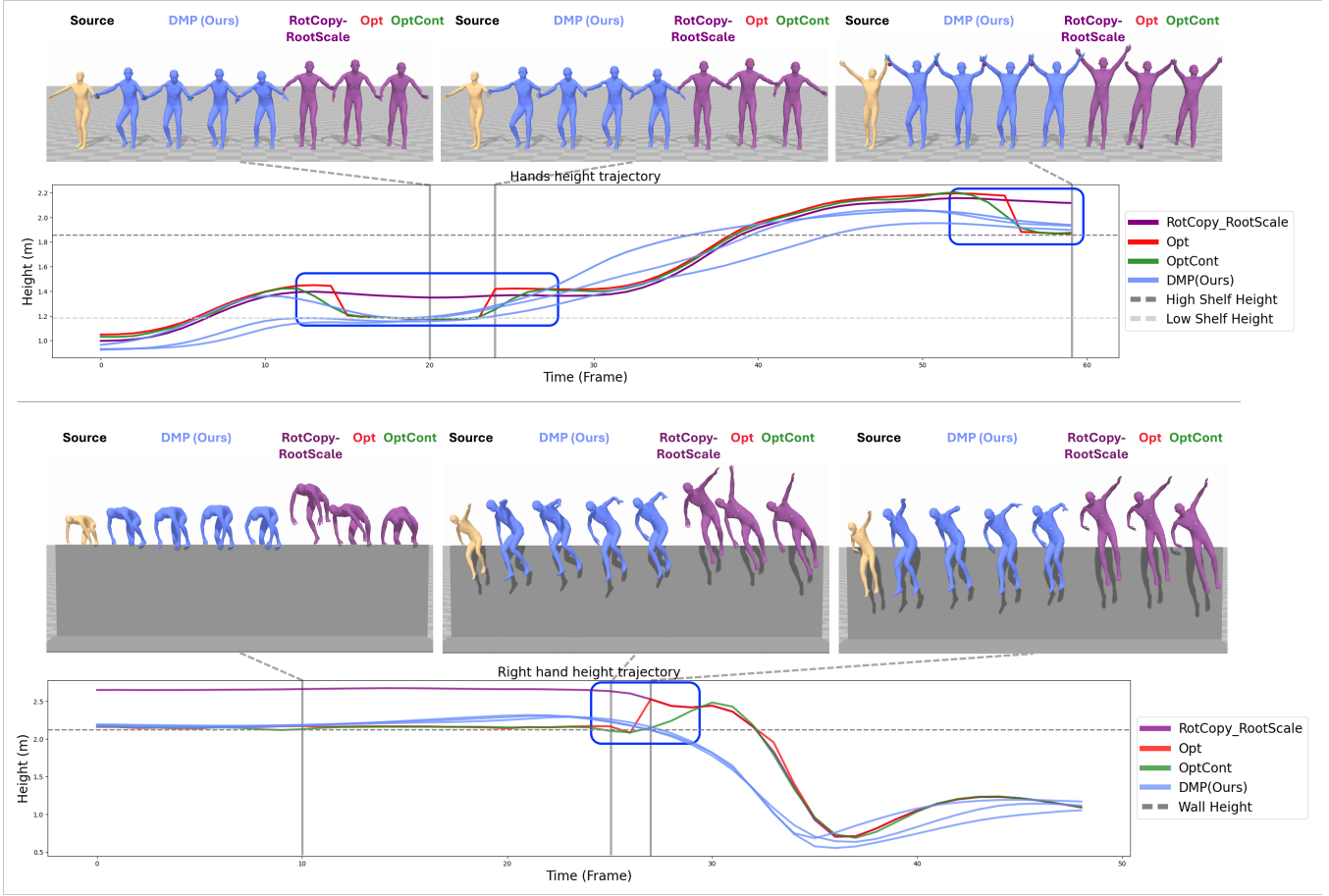


Fig. 8. Comparison of constrained retargeting on shelf-reaching (Top) and wall-jump (Bottom) motions. (Top) Positional constraints are applied to both hands when picking up and putting down an object. The plot shows the averaged height trajectory of both hands. (Bottom) Positional constraints are applied to the right hand when vaulting over the wall. The plot shows the height trajectory of the right hand. In both examples, *RotCopy-RootScale* fails to preserve hand positions. *Opt* satisfies constraints only at the specific frames where they are imposed, which results in sharp motion discontinuities around the constrained frames. *OptCont* shows smoother transitions, yet the lack of motion priors leads to awkward ease-in and ease-out motions around the constrained frames. Discontinuous motions and unnatural smoothing are highlighted in blue boxes. DMP (Ours) generates plausible retargeted motions that accurately satisfy constraints while preserving natural behaviors across the entirety of the motion sequence.

shown in Table 2. *Inference w/o Mask* and *Inference w/o Guide* disable masked inpainting and guidance during generation, respectively. While both inpainting and guidance individually can be used to incorporate constraints, applying both yields the best performance. *Train w/o Mask* corresponds to a model trained without masked inpainting, which notably degrades constraint-following performance, since the model has to rely entirely on guidance. To show the importance of our structured mask sampling strategy, *Train w/ Random Mask* trains a model using random masks, where each feature is independently masked with equal probability at every frame. This masking strategy can misallocate model capacity to learning constraint patterns that are not practically useful, leading to higher errors on the specific set of evaluation constraints. *Train w/o ConLoss* corresponds to a model trained without constraint loss, i.e., $w_{\text{con}} = 0$ in Eq. (6). While the reconstruction loss penalizes

errors for all features, the constraint loss serves to emphasize the importance of constraint matching. The ablation shows slight improvement on constraint violation metrics, except for the Root pos error. This component has relatively less impact than other design choices and could be considered optional.

7.4 Quantitative Results: Unconstrained Retargeting

Exploring the diverse space of possible retargeting solutions can be an important first step in the user workflow before introducing any detailed constraints to refine the output motion. Therefore, we also evaluate the retargeting performance of our model when no constraints are specified. For this experiment, the ground-truth target skeletons from the test split (1000 motions) are used so that retargeting results can be compared to the ground-truth target motion. Table 3 shows metrics that measure the error between the generated

Table 3. Retargeting errors and motion quality evaluation for unconstrained retargeting. Compared to heuristics-based retargeting techniques, such as rotation copy and inverse kinematics (IK), our model is able to learn motion priors from the motion paraphrasing dataset, resulting in more natural and accurate retargeting results. Recent neural retargeting method SAME suffers large errors due to overly smoothed output motions.

Method	Error to Ground Truth				Motion Quality			
	Root Pos	Heading Orient	Joint Pos	Joint Rot	FID	Foot Skate	Jerk	Penetration
RotCopy-RootScale	17.620	8.885	21.766	11.572	1.491	0.345	0.053	0.021
IK	19.003	8.893	23.464	12.244	3.511	0.341	0.063	0.036
GMR [Araujo et al. 2025]	17.752	9.287	20.720	14.210	1.501	0.327	0.050	0.056
SAME [Lee et al. 2023]	21.812	9.326	21.812	11.243	6.281	0.236	0.053	0.0003
DMP (Ours)	7.370	3.715	11.279	8.033	1.997	0.306	0.081	0.011

Table 4. Diversity (σ_p : joint positions, σ_r : joint rotations) and constraint violation (E_p : joint position, E_r : joint rotations) of retargeting results under different number of constraints. Diversity is measured as the standard deviation across generated samples and constraint violation indicates deviation from specified keyframes.

Constraints	Diversity		Constraint Violation	
	σ_p	σ_r	E_p	E_r
Unconditional	5.34	3.01	–	–
5 keyframes	4.90	2.98	9.05	4.10
10 keyframes	4.69	2.95	8.87	4.35
20 keyframes	4.59	2.91	8.40	4.40
40 keyframes	2.83	2.78	8.23	4.14

target motion and ground truth. Positions are in cm in the global coordinate frame and rotations in degrees. Motion quality metrics on the right side are implemented according to Sec 7.3.

Table 3 compares our model to multiple baseline methods. *RotCopy-RootScale* is the same method described in Sec 7.3. Simply copying rotations exhibits high distributional similarity to ground truth (lowest FID), however output motions have high errors with respect to the ground-truth target motions as copying rotations does not account for the difference in limb proportions between source and target skeletons. Since unconditional retargeting does not involve explicit keyframe constraints, *Opt* and *OptCont* cannot be applied. To ensure fair comparison, we instead include *IK* as a baseline for unconditional retargeting. The *IK* method adds an extra inverse kinematics optimization step on top of *RotCopy-RootScale* to match the head, hands, and toes of the source motion. While IK can match end-effector positions from the source motion, this does not ensure an accurate motion on the target skeleton, since joint positions can differ greatly with varying skeleton proportions, leading to large errors. *GMR* [Araujo et al. 2025] is a recent optimization-based method that retargets diverse human motion data to humanoid robots. As our source and target character shares the same skeletal topology, we treat all body joints as key joints. *GMR* achieves similar constraint violation metrics with *IK*, while *GMR* exhibits improved FID through non-uniform local scaling, which better accounts for scale differences than vanilla IK. The final baseline, *SAME* [Lee et al. 2023], is a recent neural retargeting model that we retrained on our motion paraphrasing dataset. We enlarged the model to roughly match the number of learnable parameters of our model to adapt it to

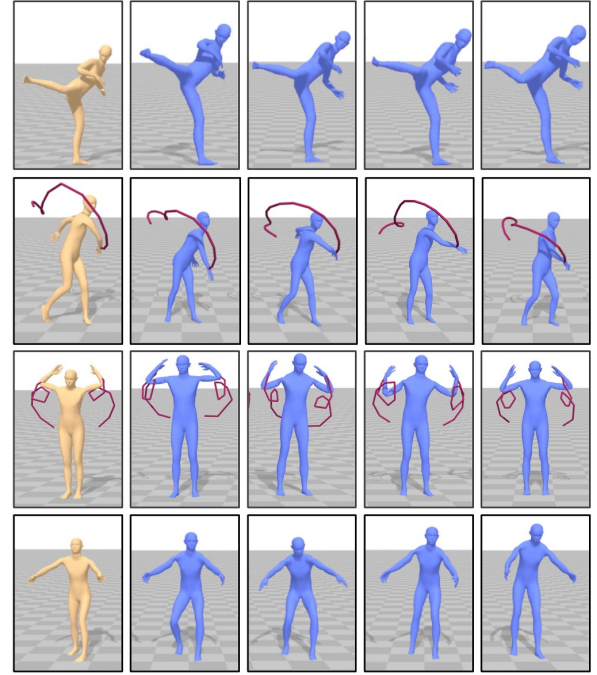


Fig. 9. DMP is able to generate diverse variations of a given motion that preserve the core semantics of the original behavior. (Kicking) Diverse variations of a kick that preserves the key kicking motion. (Throwing) Variations of a throwing motion with different arm trajectories. (Locking Dance) The stylistic wrist roll is faithfully preserved across the generated motions. (Picking up) Natural reaching motions with adjustments to for different character morphologies.

the large training dataset. *SAME* suffers from large errors to ground truth especially on root translation and joint positions, since it is a per-frame model that can accumulate global translational errors, while our model is sequence-based. This also lead to *SAME* showing large joint position errors which are also measured in the global frame. As a deterministic model, *SAME* is not able to capture the distribution of possible retargeted motions, and tends to learn the average of all motion paraphrases, making it prone to oversmoothing and loss of fine details from the source motion, as shown in the supplementary video. Overall, DMP (Ours) achieves the lowest errors to ground truth across all metrics, while also maintaining competitive motion quality in unconstrained retargeting.

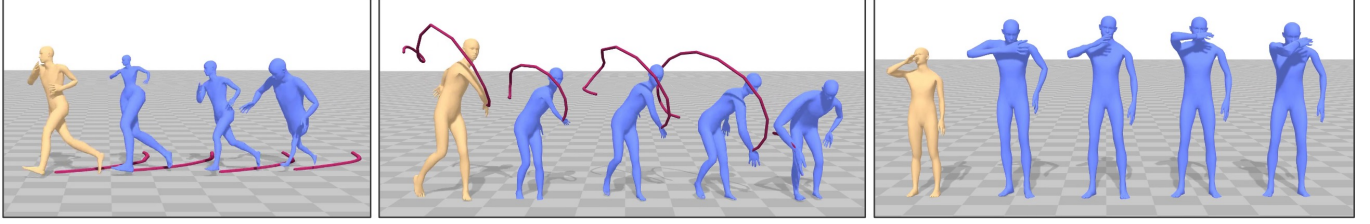


Fig. 10. Stress testing DMP on highly exaggerated body proportions. The yellow character is the source motion and blue characters indicate the paraphrased results. (Left) Running motion evaluated on characters with upper-to-lower body proportions scaled by $0.5\times$, $1.0\times$ (normal), and $2.0\times$. The model maintains circular run while adjusting its stride. (Middle) Throwing motion evaluated on characters with right arm scaled by $0.7\times$, $1.0\times$ (normal), $1.3\times$, and $1.6\times$. The model preserves the throwing semantics for deviations up to 30% ($1.3\times$), while it fails at the extreme $1.6\times$ threshold. (Right) Drinking motion with elongated forearms ($1.5\times$), normal character, and elongated upper arms ($1.5\times$, $2.0\times$). By applying guidance to preserve the relative joint position between head and the right hand, the model ensures spatial coordination despite significant imbalanced arm proportion.

7.5 Motion Paraphrasing and Semantics

We define motion semantics as the invariant characteristics of a motion that allow humans to perceive variations of execution as the same motion. This is challenging to formalize as a rigid mathematical metric because the allowable variations depend heavily on the motion and the character's context. Instead, we believe human reenactment data serves as the gold standard for defining these semantics.

By training on our motion paraphrasing dataset, DMP learns the implicit distribution of natural motion variations as shown in Figure 9. For example, with a kicking motion, while exact pose at every frame may vary, all paraphrased motions preserve the core kicking movement. When throwing a ball, the overall arm movements faithfully preserve the salient semantics of the throwing motion. Similarly, for a locking dance sequence, the model faithfully maintains distinctive circular trajectory of a wrist roll, capturing the stylistic signature that defines the dance move. Finally, when picking up an object from a shelf, a taller character might bend its knees or reach further down, providing different realizations of the same behavior that adapt to the character's morphology, while conveying the same core semantics.

7.6 Effect of Constraints on Diversity

A key property of motion paraphrasing is the ability to generate diverse retargeting solutions while gradually refining the generated motion to satisfy user-specified constraints. To quantitatively analyze this capability, we measure how diversity changes with increasing number of constraints. We generate 20 samples each for 1000 test cases, with four constraint setup (unconditional, 5, 10, 20 and 40 keyframes) and report diversity of the generation. As in Section 7.3, keyframe constraint types are chosen among root position, root orientation, joint positions, and joint rotations and constraint values are determined by the input source motions. Target skeletons are randomly sampled with a height difference of at least 15cm with respect to the source character. Diversity is quantified as the standard deviation of joint positions (σ_p) and rotations (σ_r) across the generated samples. Position metrics are in cm and rotation metrics are in degree. As more constraints are imposed, diversity is gradually reduced while the motions remain faithful to the conditions (Table 4). This highlights the flexibility of our generative model, in

contrast to baseline methods (*Opt*, *OptCont*, *IK*, *SAME*) that yield deterministic solution without diversity.

7.7 Non-Human-Like Characters

To explore the generalization performance of DMP, we evaluated our model on characters with extreme, non-human-like proportions (Figure 10). The model proves robust to even significantly non-human-like body proportions that fall outside training distribution. As illustrated in the jogging example (Figure 10, left), DMP maintains stability across a wide range of upper-to-lower body ratios ($0.5\times$ to $2.0\times$). The character with shorter legs adjusts stride length by stretching the leg more, while the character with longer leg increases knee flexion. In the throwing example (Figure 10, middle), the model successfully paraphrases the motion for characters with arm length variations up to 30%. However, the model breaks in extreme out-of-distribution cases, such as when an arm is elongated by $1.6\times$. The drinking motion (Figure 10, right) shows how constraints can complement the model's operational range to extreme proportions. With extreme imbalances of upper arm and fore arm length, the paraphrased motion approximately resembles arm lifting trajectory but fails to precisely ensure hand-to-mouth contact. By imposing relative joint position constraints, the model restores the core semantics of drinking. As our approach is data-driven, it is naturally subject to the learned distribution of training dataset. Extending these generalization boundaries to encompass even more extreme body proportions will be a promising future research.

7.8 Post Processing

While DMP achieves plausible motion retargeting under user constraints, minor artifacts such as foot skating or slight constraint deviations can occasionally occur. To resolve this, a simple post-processing can be further applied. As DMP inherently generates natural movements that closely adhere to the given constraints, standard post-processing can seamlessly eliminate residual artifacts without extensive parameter tuning. We demonstrate the ease of this refinement using the lightweight post-processing method of Kimodo [Rempe et al. 2026]. Please refer to the supplementary video for the animated results. For highly dynamic actions like spin-kicks, wall climbing, and box jumps, where maintaining perfect contact is challenging, post-processing effectively stabilizes planted feet,

Table 5. Constraint violation and motion quality evaluation for constrained retargeting with BML-MP dataset. (Top) DMP (Ours) achieves both reasonable constraint satisfaction and motion quality compared to baselines. (Bottom) Ablation studies demonstrate that incorporating both masked inpainting and guidance at inference time is necessary to achieve accurate constraint satisfaction.

Method	Constraint Violation				Motion Quality			
	Root Pos	Heading Orient	Joint Pos	Joint Rot	FID	Foot Skate	Jerk	Penetration
RotCopy-RootScale	7.308	0.000	15.477	0.000	1.221	0.311	0.050	0.006
Opt	0.067	0.001	3.881	0.001	79.383	0.463	0.373	0.012
OptCont	0.052	0.007	4.642	0.008	49.856	0.466	0.108	0.029
DMP (Ours)	3.731	1.507	12.769	4.633	2.791	0.224	0.049	0.002
Inference w/o Mask	4.265	1.399	13.927	4.930	2.850	0.225	0.047	0.002
Inference w/o Guide	14.741	6.594	22.447	11.711	1.990	0.215	0.043	0.002

Table 6. Retargeting errors and motion quality evaluation for unconstrained retargeting with BML-MP dataset. DMP (Ours) achieves the lowest error to ground-truth while maintaining reasonable motion quality.

Method	Error to Ground Truth				Motion Quality			
	Root Pos	Heading Orient	Joint Pos	Joint Rot	FID	Foot Skate	Jerk	Penetration
RotCopy-RootScale	22.579	13.453	25.703	14.816	0.932	0.296	0.055	0.006
IK	26.770	13.501	28.181	15.831	8.709	0.328	0.067	0.047
SAME [Lee et al. 2023]	26.528	12.953	26.491	17.914	2.497	1.755	0.087	0.002
DMP (Ours)	17.396	9.923	17.980	14.691	2.446	0.231	0.041	0.002

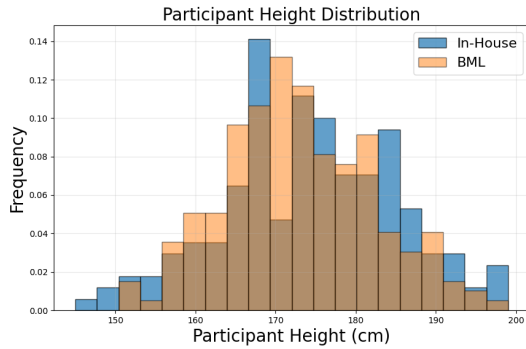


Fig. 11. Subject height distribution of in-house dataset and BML dataset (BMLmovi [Ghorbani et al. 2021] and BMLrub [Troje 2002]).

cleans up foot skating, and resolves minor floating issues. Furthermore, it improves exact constraint satisfaction, such as minimizing positional errors in a box-moving task and effectively eliminating rotational deviations in a dance example. These results indicate that our model’s outputs can be easily optimized to meet high-quality animation standards.

7.9 Evaluation with BML-MP Dataset

In addition to experiments on our proprietary in-house dataset, we evaluate our framework using a public dataset. We construct BML-MP, a motion paraphrasing dataset using BMLmovi [Ghorbani et al. 2021] and BMLrub [Troje 2002] datasets. BMLmovi and BMLrub comprise of 197 subjects performing similar motions across 37 categories (totaling 11.6 hours of motion data), and are therefore suitable for constructing a motion paraphrasing dataset. The participants

include 118 females and 79 males, with heights ranging from 152cm to 199cm, roughly following a normal distribution with a mean of 174cm and a standard deviation of 9cm (Fig. 11).

The dataset includes everyday actions and sports movements: kick, point, clap, cross arms, run in place, crawl, walk, wave, check watch, side gallop, vertical jump, sit chair, sit ground, throw and catch, run, scratch head, phone call, jumping jack, take a photo, walk in a treadmill (four speed variations: fast, jog, normal, and slow), motorcycle, walk, jog, sit, lift objects (two variations: light and heavy object), throw hard, kick, jump, knock, and walk in circle. The datasets also include freeform improvisation for each participant, but we discard these as they are inappropriate for finding motion paraphrasing pairs. The raw motion captures vary from 1 to 133 seconds in length, with an average of 8 seconds.

We exclude 3 motion categories (take a photo, pick up a heavy object, walk in circle) and 20 subjects (10 male and 10 female subjects) from the training set, using them as test motions and test subjects, respectively. The resulting BML-MP contains 2M motion paraphrasing pairs, each 1 to 4 seconds long. During training, random scaling of the character skeletons is applied, as in our in-house dataset. We train and evaluate constrained and unconstrained retargeting with BML-MP with the same setup as Sec 7.3 and Sec 7.4.

Overall, the results exhibit similar trends as the evaluation using our in-house motion paraphrasing dataset. When given constraints (Table 5), Optimization-based methods (*Opt*, *OptCont*) results in large discontinuity and lower motion quality despite more accurate constraint satisfaction, while DMP achieves both reasonable constraint satisfaction and motion quality. The ablations confirm that applying both the mask and guidance is crucial for accurate constraint satisfaction. In unconditional experiments (Table 6), DMP achieves the best results compared to both the optimization-based

method (*IK*) and previous learning-based method (*SAME* [Lee et al. 2023]). While our proprietary dataset comprises hundreds of hours of motions across thousands of categories, BML-MP is limited to about 10 hours of motion data with fewer than 40 motion categories. Due to this limitation, the model results in higher quantitative errors and sometimes struggles to generalize to new motions, as shown in the supplementary video.

8 Discussion

We have introduced Directable Motion Paraphrasing (DMP), a generative retargeting approach that captures the distribution of possible retargeting solutions while enabling user directability through kinematic constraints. As shown in Sec 7, our approach learns a prior of human motions that captures both variations and invariances through motion paraphrases, leading to improved motion quality and semantic preservation over heuristics-based retargeting methods. Moreover, our diffusion architecture offers users control over retargeting outputs through masked inpainting and diffusion guidance with kinematic constraints. While DMP shows promising results, in some cases it can fail to preserve semantically salient characteristics of the source motion, and our model also does not always preserve high-frequency motion details in the original source motion. However, this can often be ameliorated through constraints, for example by conserving relative distances between important body parts or constraining joint rotations to closely capture the details of the source motion. While our system is able to accommodate a wide range of body proportions due to the large motion paraphrasing dataset, it is currently limited to a single skeleton topology. An exciting direction for future work would be to explore motion paraphrasing between characters with different topologies and body meshes.

Acknowledgments

We thank Umar Iqbal, Ye Yuan, Mathis Petrovich, Jiashun Wang, and Purvi Goel for insightful discussions. Jungdam Won is partially supported by the NRF of Korea under Grant No. RS-2024-00450647 and by the IITP under Grant Nos. RS-2025-25442338 (AI Star Fellowship Support Program, Seoul National Univ.), RS-2021-II211343 (AI Graduate School Program, Seoul National Univ.), and IITP-2026-RS-2020-II201460 (Information Technology Research Center).

References

- Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. 2020. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 62–1.
- Rami Ali Al-Asghar, Taku Komura, and Myung Geol Choi. 2013. Relationship descriptors for interactive motion adaptation. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 45–53.
- Mazen Al Borno, Ludovic Righetti, Michael J Black, Scott L Delp, Eugene Fiume, and Javier Romero. 2018. Robust Physics-based Motion Retargeting with Realistic Body Shapes. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 81–92.
- Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–20.
- Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. Gesturediffclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–18.
- Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. 2025. Retargeting Matters: General Motion Retargeting for Humanoid Motion Tracking. *arXiv preprint arXiv:2510.02252* (2025).
- Nikos Athanasiou, Alpar Ceske, Markos Diomatari, Michael J Black, and Gül Varol. 2024. MotionFix: Text-driven 3d human motion editing. *arXiv preprint arXiv:2408.00712* (2024).
- Emad Barsoum, John Kender, and Zicheng Liu. 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1418–1427.
- Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. 2024. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*. 1–10.
- Théo Cheynel, Thomas Rossi, Baptiste Bellot-Gurlet, Damien Rohmer, and Marie-Paule Cani. 2025. ReConForM: Real-time Contact-aware Motion Retargeting for more Diverse Character Morphologies. In *Computer Graphics Forum*. Wiley Online Library, e70028.
- Kwang-Jin Choi and Hyeon-Seok Ko. 1999. On-line motion retargeting. In *Proceedings of Pacific Graphics*. 32–42.
- Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. 2023. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=OnD9zGAGT0k>
- Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. 2024. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*. 1–9.
- Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023a. Mofusion: A Framework for Denoising-Diffusion-Based Motion Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9760–9770.
- Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023b. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9760–9770.
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- William Falcon and The PyTorch Lightning team. 2019. *PyTorch Lightning*. doi:10.5281/zenodo.3828935
- Reuben Feinman. 2021. Pytorch-minimize: a library for numerical optimization with autograd. <https://github.com/rfeinman/pytorch-minimize>
- Saeed Ghorbani, Kimia Mahdavian, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. 2021. MoVi: A large multi-purpose human motion and video dataset. *Plos one* 16, 6 (2021), e0253157.
- Michael Gleicher. 1998. Retargeting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques (I3D)*. 33–42.
- Deepak Gopinath and Jungdam Won. 2020. fairmotion - Tools to load, process and visualize motion capture data. Github. <https://github.com/facebookresearch/fairmotion>
- Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.
- Edmond SI Ho, Taku Komura, and Chiew-Lan Tai. 2010. Spatial relationship preserving character motion adaptation. In *ACM SIGGRAPH 2010 papers*. 1–8.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.
- Lei Hu, Zihao Zhang, Yongjing Ye, Yiwen Xu, and Shihong Xia. 2024. Diffusion-based Human Motion Style Transfer with Semantic Guidance. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15169.
- Lei Hu, Zihao Zhang, Chongyang Zhong, Boyuan Jiang, and Shihong Xia. 2023. Pose-aware attention network for flexible motion retargeting by body part. *arXiv preprint arXiv:2306.08006* (2023).
- Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16750–16761.
- Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2151–2162.
- Manuel Kaufmann, Velko Vechev, and Dario Mylonopoulos. 2022. *aitviewer*. doi:10.5281/zenodo.10013305
- Jehee Lee and Sung Yong Shin. 1999. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (I3D)*. 39–48.
- Sunmin Lee, Taehong Kang, Jungnam Park, Jehee Lee, and Jungdam Won. 2023. SAME: Skeleton-Agnostic Motion Embedding for Character Animation. In *SIGGRAPH Asia*

- 2023 *Conference Papers*. 1–11.
- Jiaman Li, Karen Liu, and Jiajun Wu. 2023. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17142–17151.
- Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. 2022. Ganimator: Neural motion synthesis from a single sequence. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–12.
- Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13401–13412.
- Jongin Lim, Hyung Jin Chang, and Jin Young Choi. 2019. PMnet: Learning of Disentangled Pose and Movement for Unsupervised Motion Retargeting.. In *Proceedings of British Machine Vision Conference (BMVC)*, Vol. 2. 7.
- Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. 2020. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 40–1.
- Zhiqiang Liu, Antonio Mucherino, Ludovic Hoyet, and Franck Multon. 2018. Surface based motion retargeting by preserving spatial relationship. In *Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 1–11.
- Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.
- Meinard Müller. 2015. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Vol. 5. Springer.
- Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 8162–8171.
- Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. 2008. FMDistance: A Fast and Effective Distance Function for Motion Capture Data. *Eurographics (Short Papers)* 7, 10 (2008).
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).
- Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10985–10995.
- Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. 2024. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1546–1555.
- Zoran Popović and Andrew Witkin. 1999. Physically based motion transformation. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. ACM Press/Addison-Wesley Publishing Co., USA, 11–20. doi:10.1145/311535.311536
- Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit Haim Bermano, and Daniel Cohen-Or. 2024. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. In *SIGGRAPH Asia 2024 Conference Papers*. 1–13.
- Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11488–11499.
- Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13756–13766.
- Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. 2026. Kimodo: Scaling Controllable Human Motion Generation. *arXiv:2603.15546* (2026).
- Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. 2023. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418* (2023).
- Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. 2024. Interactive character control with auto-regressive motion diffusion models. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–14.
- Hyun Joon Shin, Jehhee Lee, Sung Yong Shin, and Michael Gleicher. 2001. Computer puppetry: An importance-based approach. *ACM Transactions on Graphics (TOG)* 20, 2 (2001), 67–94.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
- Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. 2024. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 821–830.
- Justin Studer, Dhruv Agrawal, Dominik Borer, Seyedmorteza Sadat, Robert W Sumner, Martin Guay, and Jakob Buhmann. 2024. Factorized motion diffusion for precise and character-agnostic motion inbetweening. In *Proceedings of the 17th ACM SIGGRAPH conference on motion, interaction, and games*. 1–10.
- Seyoon Tak and Hyeon-Seok Ko. 2005. A physically-based motion retargeting filter. *ACM Transactions on Graphics (TOG)* 24, 1 (2005), 98–117.
- Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=SJ1kSyO2jwu>
- Shashank Tripathi, Omid Taheri, Christoph Lassner, Michael Black, Daniel Holden, and Carsten Stoll. 2025. HUMOS: Human Motion Model Conditioned on Body Shape. In *European Conference on Computer Vision*. Springer, 133–152.
- Nikolaus F Troje. 2002. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision* 2, 5 (2002), 2–2.
- Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. 2021. Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9720–9729.
- Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. 2018. Neural kinematic networks for unsupervised motion retargeting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8639–8648.
- Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2023. Omni-control: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580* (2023).
- Yuan Xu, Xiaoxuan Ma, Jiajun Su, Wentao Zhu, Yu Qiao, and Yizhou Wang. 2024. Score-hypo: Probabilistic human mesh estimation with hypothesis scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 979–989.
- Xiaohang Yang, Qing Wang, Jiahao Yang, Gregory Slabaugh, and Shanxin Yuan. 2025. STaR: Seamless Spatial-Temporal Aware Motion Retargeting with Penetration and Consistency Constraints. *arXiv preprint arXiv:2504.06504* (2025).
- Zijie Ye, Jia-Wei Liu, Jia Jia, Shikun Sun, and Mike Zheng Shou. 2024. Skinned motion retargeting with dense geometric interaction perception. *Advances in Neural Information Processing Systems* 37 (2024), 125907–125934.
- Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. 2024. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*. Springer, 246–263.
- Ye Yuan and Kris Kitani. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*. Springer, 346–364.
- Haodong Zhang, Zhike Chen, Haocheng Xu, Lei Hao, Xiaofei Wu, Songcen Xu, Rong Xiong, and Yue Wang. 2024a. Unified cross-structural motion retargeting for humanoid characters. *IEEE Transactions on Visualization and Computer Graphics* (2024).
- Haodong Zhang, Zhike Chen, Haocheng Xu, Lei Hao, Xiaofei Wu, Songcen Xu, Zhen-song Zhang, Yue Wang, and Rong Xiong. 2024b. Semantics-aware motion retargeting with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2155–2164.
- Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Yue Wang, and Zhigang Tu. 2023a. Skinned motion retargeting with residual perception of motion semantics & geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13864–13872.
- Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023b. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14730–14740.
- Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022).
- Zeyu Zhang, Yiran Wang, Wei Mao, Danning Li, Rui Zhao, Biao Wu, Zirui Song, Bohan Zhuang, Ian Reid, and Richard Hartley. 2025. Motion anything: Any to motion generation. *arXiv preprint arXiv:2503.06955* (2025).
- Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2025. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*. Springer, 405–421.
- Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5745–5753.