Dynamic Motion Blending for Versatile Motion Editing

Nan Jiang[1,2*] Hongjie Li[1*] Ziye Yuan[1*] Zimo He[1,2,3]
Yixin Chen[2] Tengyu Liu[2] Yixin Zhu[1✉] Siyuan Huang[2✉]

[1] Institute for AI, Peking University [2] State Key Laboratory of General Artificial Intelligence, BIGAI
[3] Yuanpei College, Peking University [*] Equal contribution ✉ yixin.zhu@pku.edu.cn, syhuang@bigai.ai

https://awfuact.github.io/motionrefit/

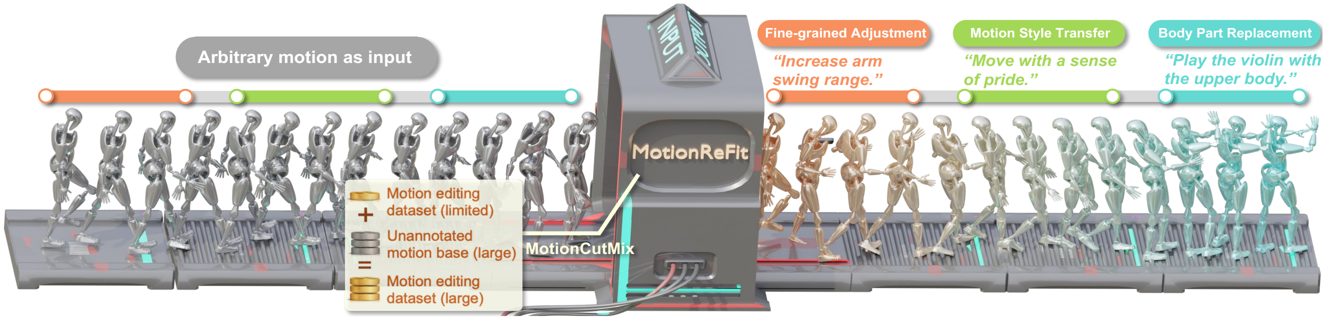


Figure 1. **MotionReFit**, a universal framework for motion editing that handles various scenarios simply from textual guidance, offering both spatial and temporal editing capabilities. MotionReFit is supercharged with our proposed **MotionCutMix** training strategy, which leverages large-scale unannotated motion databases to augment the scarce motion editing triplets, enabling robust and generalizable editing.

Abstract

Text-guided motion editing enables high-level semantic control and iterative modifications beyond traditional keyframe animation. Existing methods rely on limited pre-collected training triplets (original motion, edited motion, and instruction), which severely hinders their versatility in diverse editing scenarios. We introduce MotionCutMix, an online data augmentation technique that dynamically generates training triplets by blending body part motions based on input text. While MotionCutMix effectively expands the training distribution, the compositional nature introduces increased randomness and potential body part incoordination. To model such a rich distribution, we present MotionReFit, an auto-regressive diffusion model with a motion coordinator. The auto-regressive architecture facilitates learning by decomposing long sequences, while the motion coordinator mitigates the artifacts of motion composition. Our method handles both spatial and temporal motion edits directly from high-level human instructions, without relying on additional specifications or Large Language Models (LLMs). Through extensive experiments, we show that MotionReFit achieves state-of-the-art performance in text-guided motion editing. Ablation studies further verify that MotionCutMix significantly improves the model's generalizability while maintaining training convergence.

1. Introduction

Text-guided motion editing has emerged as a fundamental task in computer vision and animation [7, 26, 70], enabling creators to perform *semantic edits* (*e.g.*, altering the right-hand movement to a circular motion) and *style edits* (*e.g.*, performing the motion in an angry style) through natural language instructions. Despite recent advances, current approaches [7, 56, 70] face three critical limitations in achieving efficient, flexible, generalizable, and natural motion editing.

First, following InstructPix2Pix [10], existing methods [6, 7] rely on fixed triplets of original motion, edited motion, and editing instructions. This dependency severely restricts their ability to generalize across diverse scenarios, especially for style edits and novel motion-instruction combinations. Second, current models require explicit specification of body parts as auxiliary information, limiting their capability to autonomously comprehend high-level semantic instructions. Third, generating edited motions with smooth spatial-temporal transitions remains challenging.

To address these limitations, we introduce **MotionCutMix**, a training technique that synthesizes novel triplets by blending body parts from multiple motion sequences. This approach leverages abundant unannotated motion data to augment expensive annotated editing triplets. Specifically, we employ a soft-mask mechanism for spatial blending of body parts, producing dynamically composited triplets of original motion, edited motion, and corresponding language instruction. This enables end-to-end editing using purely natural language input.

However, training with MotionCutMix introduces two potential side-effects in motion generation: increased randomness and body part incoordination. To address these issues, we propose **MotionReFit** (Motion REgeneration From Input Text), an auto-regressive conditional diffusion model accompanied by a motion coordinator, as Fig. 1 shows. By employing an auto-regressive strategy, the motion is generated segment by segment, significantly facilitating convergence during training by decomposing long sequences. This approach also enables temporal editing with a smooth transition. To mitigate the incoordination in generated motion, we train a motion coordinator as a discriminator to assess whether a motion segment is the result of composition. This discriminator is used to refine the diffusion process as guidance, encouraging the generated motion segments to adherently resemble the pattern of original motions and avoiding model collapses to unnatural mode.

We extensively evaluate our approach using our proposed **STANCE** (Style Transfer, Fine-Grained Adjustment, and Body Part Replacement) dataset, which is developed for three text-guided motion editing tasks. Our experimental evaluations demonstrate that MotionReFit achieves high-fidelity edits across all three tasks while faithfully following the provided textual instructions. Through comprehensive ablation studies, we find that incorporating MotionCutMix substantially enhances the model's generalization capability, particularly when training data is limited. Importantly, despite augmenting training data complexity, MotionCutMix does not significantly impact the training convergence efficiency, allowing the model to benefit from expanded motion diversity without computational overhead.

Our primary contributions are threefold:

- We present MotionReFit, the first universal text-guided motion editing framework that achieves unrestricted editing capabilities for both body parts and temporal sequences. Powered by segmental motion synthesis mechanism and attention-based local-global refinement strategy, MotionReFit requires only original motion and editing instruction as input while delivering superior instruction adherence and motion naturalness.
- We introduce MotionCutMix, a dynamic training technique that augments motion editing triplets online, enabling robust generalization, even with limited annotated data.
- We contribute MotionCutMix, a motion-captured and manually annotated dataset for three editing tasks: body part replacement, fine-grained adjustment, and motion style transfer, providing diverse and high-quality examples for training and evaluation.

2. Related Work

Data-Driven Motion Generation With access to large-scale motion datasets [18, 40, 46, 50], early motion generation approaches focused on predicting future motion [3, 66]. Recent efforts have incorporated action labels and language descriptions to enhance the relevance and specificity of generated motions [8, 19, 23, 35, 42, 55, 61, 72]. The emergence of diffusion models [21, 52] has marked a significant advancement in motion synthesis [12–14, 29, 31, 53, 56, 65, 69, 70]. Several approaches [56, 64, 69, 70] have introduced motion editing capabilities. MDM [56] supports part-level motion inpainting and temporal inbetweening, while FineMoGen [70] leverages LLMs to interpret and execute editing instructions. However, these methods fail to simultaneously handle semantic and style edits.

Motion Style Transfer Early approaches in style transfer primarily relied on handcrafted features to address the complexities of defining and manipulating motion styles [4, 58, 62]. With the advent of deep learning, contemporary studies have favored data-driven techniques that leverage large datasets to extract and learn style features, utilizing approaches such as GAN [15], AdaIN [2], and Diffusion [11, 47, 67]. While some methods employ neural networks trained on explicit pairs of original and edited motion styles [9, 24, 27, 39, 60, 63] to directly translate specific movement patterns, others explore unpaired training strategies [2, 11, 25, 28, 54] to infer style from unaligned motion data or video inputs. However, despite these advancements in style transfer techniques, current methodologies predominantly address non-semantic motions and remain limited in their capacity to tailor arbitrary motions based on specific semantic textual descriptions.

Motion Editing Motion editing, while sharing similarities with motion style transfer, remains comparatively under-explored. Early research focused on specific motion attributes such as adjusting motion paths [16, 33, 36], adapting motions to different skeletal structures [1], or altering motion-induced emotions [58].

In terms of semantic editing, Tevet et al. [55] and Holden et al. [22] proposed embedding motion sequences into latent vectors that encapsulate semantic information. However, this approach faces fundamental challenges as the embeddings may lack the fine-grained detail necessary for precise editing, and the latent space may not be sufficiently disentangled. Recent diffusion-based approaches [32, 45, 56, 69] have enabled editing of existing motions through inpainting conditioned on textual instructions. However, these methods fix the joints of the remaining body parts, requiring clear delineation of the parts to be edited.

Another significant line of research facilitates editing through motion composition, including temporal composition [5, 49, 51, 57], spatial composition [6, 42], and com-

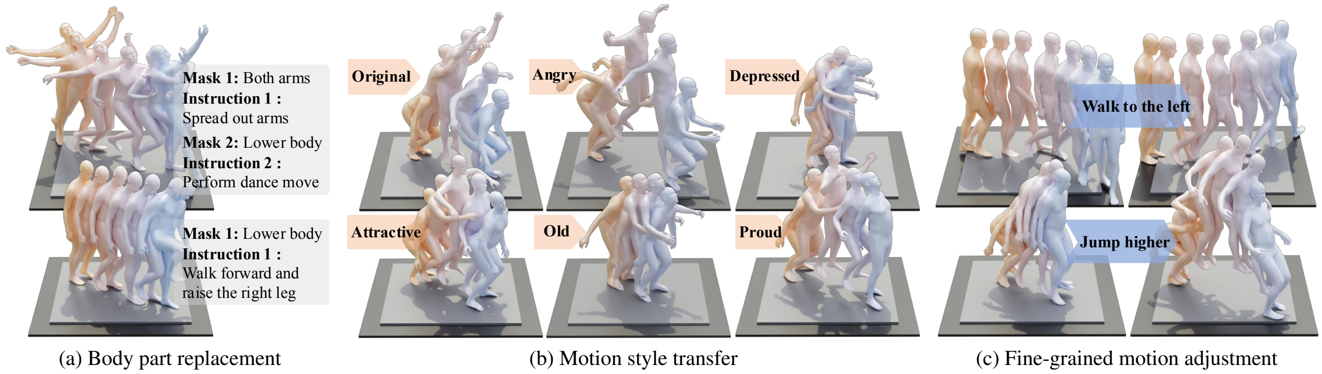


Figure 2. **Sample sequences from our STANCE dataset.** Our work introduces three complementary datasets: (a) a body part replacement dataset comprising 13,000 sequences from HumanML3D [18], annotated with an average of 2.1 body masks and corresponding motion descriptions; (b) a motion style transfer dataset containing 2 hours of new MoCap recordings that recreate HumanML3D sequences in various styles; and (c) a fine-grained motion adjustment dataset featuring 16,000 annotated triplets of generated motion pairs with their corresponding descriptions.

prehensive timeline control frameworks [44]. Recent works such as FineMoGen [70], Iterative Motion Editing [17], and COMO [26] leverage foundation models for generating and editing motion, but they fail to handle arbitrary motion inputs without annotation. The work most similar to ours is TMED [7], which employs a conditional diffusion model using both original motion and instructions as inputs, without requiring additional data. However, TMED's training on a limited set of triplets (original, edited, and instruction) hinders its generalizability to broader compositions, and it does not effectively handle temporal composition.

Addressing these limitations, our method provides an end-to-end solution that does not require additional user inputs while effectively handling a diverse range of motion-instruction compositions with the capability for both spatial and temporal edits.

3. Problem Formulation and Representations

Text-Guided Motion Editing Given an original motion sequence $\mathcal{M}_{\text{ori}}$ and an editing instruction $\mathcal{E}$ that specifies desired modifications, our goal is to generate an edited motion sequence $\mathcal{M}_{\text{edit}}$ that satisfies the following objectives:

- $\mathcal{M}_{\text{edit}}$ should faithfully implement the modifications specified by $\mathcal{E}$, such as changes in motion style, intent, or specific body part movements.
- $\mathcal{M}_{\text{edit}}$ should maintain the integrity of $\mathcal{M}_{\text{ori}}$ by preserving aspects not explicitly specified by $\mathcal{E}$.

Human Motion Representations Our approach employs two complementary representations derived from the SMPL-X model [41]. For direct motion manipulation, we use a keypoint-based representation $\mathcal{M}^{\mathcal{K}} \in \mathbb{R}^{L\times N_K \times 3}$, where L denotes sequence length and $N_K = 28$ represents the number of keypoints. These keypoints comprise 22 primary body joints from SMPL-X, supplemented by four finger joints (ring and index fingertips of both hands) for wrist pose determination, and two additional head joints to capture detailed head movements. In this representation, hands are treated as rigid bodies without detailed finger articulation. For compatibility with standard motion frameworks, we also utilize the SMPL-X parameter-based representation $\mathcal{M}^{\mathcal{S}} = \{\mathbf{t}, \boldsymbol{\phi}, \mathbf{r}\}$. This representation consists of root translation $\mathbf{t} \in \mathbb{R}^{L\times 3}$, global orientation $\boldsymbol{\phi} \in \mathbb{R}^{L\times 3}$, and body pose parameters $\mathbf{r} \in \mathbb{R}^{L\times N_J\times 3}$, where $N_J = 21$ aligns with SMPL-X formulations. We use the mean body shape by setting $\boldsymbol{\beta}$ to zero.

These representations are interconvertible: Forward Kinematics transforms $\mathcal{M}^{\mathcal{S}}$ to $\mathcal{M}^{\mathcal{K}}$, while the reverse mapping uses a lightweight neural network followed by optimization to obtain $\mathcal{M}^{\mathcal{S}}$ from $\mathcal{M}^{\mathcal{K}}$. For simplicity, we omit representation superscripts when discussing motion in general terms. Details of motion representations and their conversions are in Appendices B.1 and B.3, respectively.

4. Training Data Construction

This section details the construction of training triplets $\{\mathcal{M}_{\text{ori}}, \mathcal{M}_{\text{edit}}, \mathcal{E}\}$. We first present our proposed STANCE dataset in Sec. 4.1. We then introduce a key motion composition operator in Sec. 4.2, followed by the rules for constructing triplets across various editing settings in Sec. 4.3.

4.1. STANCE Dataset

Our STANCE dataset introduces three specialized components targeting common editing scenarios, as shown in Fig. 2. Each component is carefully curated and verified by trained human annotators. Additional details for our STANCE dataset are available in Appendix D.

Body Part Replacement This editing type focuses on semantic edits where specific body parts are modified according to text instructions while preserving the motion of other parts. We improve upon previous approaches like [6] that relied on LLMs by having human annotators analyze rendered motions from the HumanML3D dataset [18] to assess body part participation. As illustrated in Fig. 2a, sequences can contain multiple mask sets, each annotated with descriptions of the masked body part's motion. We also introduce soft masks, detailed in Sec. 4.2, to enable spatial blending.

Style Transfer As a type of style edit, this category aims to modify motion style without altering semantic content based on language instructions. We address the general case of style transfer across both locomotion and semantic motions. To overcome the lack of paired motions with identical semantics but different styles, we created a new MoCap dataset using the Vicon system. Professional actors recreated HumanML3D sequences in various styles (*e.g.*, old, proud, depressed), resulting in 2 hours of motion comprising 750 stylized sequences.

Fine-grained Motion Adjustment This type of style edit enables detailed modifications without semantic changes (*e.g.*, "raise the right arm higher"). We introduce a novel approach that improves upon previous works like MotionFix [7], which relied on TMR [43] representations for motion pairing. Instead, we utilize MLD [12] as a text-to-motion generator to create 16 variants per instruction by perturbing the motion latent space. These variants are paired one-to-one, with human annotators describing the required transformations between pairs. After filtering out unnatural motions and unclear descriptions, we obtain 16,000 high-quality annotated triplets.

4.2. Spatial Motion Blending

As illustrated in Fig. 3, spatial motion blending enables the synthesis of novel motions by combining selected body parts from a source motion $\mathcal{M}_{\text{src}}$ with a target motion $\mathcal{M}_{\text{tgt}}$, guided by annotated masks. A mask is defined as $\mathbf{M} \subseteq \{0, 1, \ldots, N_j\}$, where $j \in \mathbf{M}$ indicates the j^{th} joint (including pelvis) is selected. The blending process is guided by two annotated masks: a hard part $\mathbf{M}_{\text{hard}}$ and a soft part $\mathbf{M}_{\text{soft}}$, ensuring $\mathbf{M}_{\text{hard}} \cap \mathbf{M}_{\text{soft}} = \varnothing$. Joints within $\mathbf{M}_{\text{hard}}$ directly inherit rotations from $\mathcal{M}_{\text{tgt}}$, while those in $\mathbf{M}_{\text{soft}}$ undergo interpolation between source and target motions, ensuring smooth spatial transitions and motion coherence.

We denote the spatial motion blending process as $\text{BLD}(\mathcal{M}_{\text{src}}, \mathcal{M}_{\text{tgt}}, \{\mathbf{M}_{\text{hard}}, \mathbf{M}_{\text{soft}}\})$. The resulting blended motion $\mathcal{M}_{\text{bld}} = \{\mathbf{t}^{\text{bld}}, \phi^{\text{bld}}, \{\mathbf{r}_j^{\text{bld}}\}_{j=1}^{N_J}\}$ is computed following these rules for each joint j:

$$\begin{cases} \mathbf{r}_j^{\text{bld}} = \mathbf{r}_j^{\text{tgt}} & \text{if } j \in \mathbf{M}_{\text{hard}} \\ \mathbf{r}_j^{\text{bld}} = \text{SLERP}(\mathbf{r}_j^{\text{src}}, \mathbf{r}_j^{\text{tgt}}, \alpha) & \text{if } j \in \mathbf{M}_{\text{soft}} \\ \mathbf{r}_j^{\text{bld}} = \mathbf{r}_j^{\text{src}} & \text{if } j \notin \mathbf{M}_{\text{hard}} \text{ and } j \notin \mathbf{M}_{\text{soft}} \end{cases}$$

where $\mathbf{r}^{\text{src}}$, $\mathbf{r}^{\text{tgt}}$, and $\mathbf{r}^{\text{bld}}$ represent joint rotations in the source, target, and blended motions respectively. The interpolation employs Spherical Linear Interpolation (SLERP) with a factor α, which is randomly varied during training to increase motion diversity.

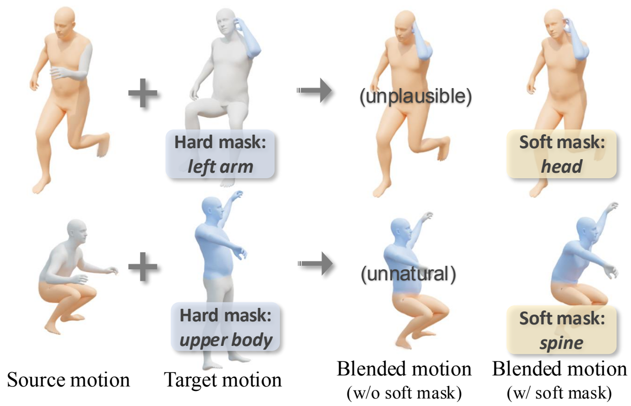


Figure 3. **Illustration of spatial motion blending.** We compare hard and soft masking approaches, showing how soft masks enable smoother transitions between body parts and eliminate unnatural artifacts at motion boundaries.

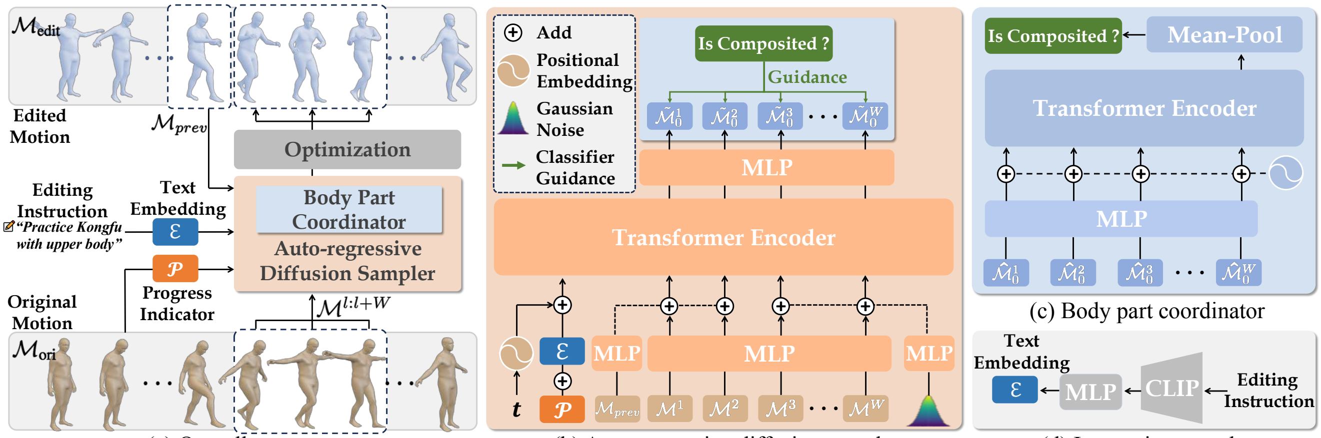
The global properties of the blended motion—orientation ϕ^{bld} and translation $\mathbf{t}^{\text{bld}}$—are determined by the lower body motion. When the pelvis is included in $\mathbf{M}_{\text{hard}}$, the root pose follows $\mathcal{M}_{\text{tgt}}$; otherwise, it inherits from $\mathcal{M}_{\text{src}}$. This approach ensures consistency between the pelvis and the dominant lower body motion.

4.3. MotionCutMix

We propose MotionCutMix, a training technique that augments the limited motion data for training by leveraging variants from a larger motion database, which can be unannotated. Inspired by image augmentation [68], MotionCutMix generates synthetic training samples through spatial motion blending on the training data. This enables the model to learn from diverse examples, capture high-level dependencies between original and edited motions, and enhance editing performance even with limited annotated training data.

MotionCutMix applies universally to both semantic and style edits, though with different composition rules. For semantic edits, MotionCutMix randomly selects $\mathcal{M}_{\text{src}}$ from the large motion base and $\mathcal{M}_{\text{tgt}}$ from the dataset with body mask annotation. The training triplet $\{\mathcal{M}_{\text{ori}}, \mathcal{M}_{\text{edit}}, \mathcal{E}\}$ consists of $\mathcal{M}_{\text{ori}} = \mathcal{M}_{\text{src}}$ and $\mathcal{M}_{\text{edit}} = \text{BLD}(\mathcal{M}_{\text{src}}, \mathcal{M}_{\text{tgt}}, \mathbf{M}_{\text{tgt}})$, where $\mathbf{M}_{\text{tgt}}$ is the body mask annotated to $\mathcal{M}_{\text{tgt}}$. The editing instruction $\mathcal{E}$ is associated with $\mathbf{M}_{\text{tgt}}$, describing how the masked body part changes from $\mathcal{M}_{\text{src}}$ to $\mathcal{M}_{\text{tgt}}$.

Style edits present a different challenge since their parts requiring edits are already paired and cannot be randomly composited. To enable the model to learn generalized editing from limited data pairs, we split the editing into lower and upper bodies. For a source-target motion pair from the annotated dataset, MotionCutMix



(a) Overall structure (b) Auto-regressive diffusion sampler (d) Instruction encoder

Figure 4. **Overview of MotionReFit.** Our auto-regressive approach processes the original motion through sliding windows, where body keypoints are encoded for input to a transformer-based motion diffusion model. To ensure motion continuity, noise is applied starting from the third frame while preserving the first two frames. The model incorporates an additional token integrating the editing instruction, diffusion step, and progress indicator. The generated keypoints undergo SMPL-X optimization and merging to create the final edited motion. To enhance body part coordination, we employ a discriminator trained to identify motion segments composed of multiple source motions, which guides the denoising process through classifier guidance.

randomly substitutes the non-edited body part of both $\mathcal{M}_{\text{src}}$ and $\mathcal{M}_{\text{tgt}}$ with the same motion sequence $\mathcal{M}_{\text{ext}}$ selected from an extra motion base. The blended pairs become $\mathcal{M}_{\text{ori}} = \text{BLD}(\mathcal{M}_{\text{ext}}, \mathcal{M}_{\text{src}}, \mathbf{M}_{\text{edited-part}})$ and $\mathcal{M}_{\text{edit}} = \text{BLD}(\mathcal{M}_{\text{ext}}, \mathcal{M}_{\text{tgt}}, \mathbf{M}_{\text{edited-part}})$, while $\mathcal{E}$ describes the style change on specific body parts.

MotionCutMix effectively creates $N_L \times N_S$ original-edited pairs from N_S annotated motion triplets, where N_L denotes the size of the large motion base. By exposing the model to diverse motion combinations, MotionCutMix enables better generalization and adherence to editing instructions.

5. MotionReFit

Our model performs end-to-end editing on arbitrary input motion by leveraging MotionCutMix for creating training triplets. As shown in Fig. 4, the framework consists of three key components: an auto-regressive motion diffusion model, a body part coordinator, and multiple condition encoders.

5.1. Motion Diffusion Model

At the core of our approach is an auto-regressive conditional diffusion model that generates edited motion segment by segment, guided by the original motion and text instruction. The model processes keypoint-based representations of human motion segments $\mathcal{M}^{l:l+W}$, where l denotes the start frame and W is the window size. For notation simplicity, we refer to $\mathcal{M}$ as "the motion in the current segment" throughout our discussion. Each segment $\mathcal{M}$ is transformed to a local coordinate system based on the root transformation of its initial frame, as detailed in Appendix B.2.

Following the Denoising Diffusion Probabilistic Models (DDPM) [21] framework, we implement a forward diffusion process as a Markov Chain that progressively adds noise to clean edited motion segments $\mathcal{M}_{\text{edit}}$ over T steps. Using $\mathcal{M}_t$ to denote the noisy version of $\mathcal{M}_{\text{edit}}$ at diffusion step t, the noise addition process follows:

$$q(\mathcal{M}_t|\mathcal{M}_{t-1}) = \mathcal{N}(\mathcal{M}_t; \sqrt{1-\beta_t}\mathcal{M}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $\beta_t \in (0,1)$ is a variance schedule controlling noise magnitude per step, and $\mathbf{I}$ is the identity matrix.

The reverse denoising process is learned by a network ϵ_θ (Appendix B.4), which sequentially denoises samples across T steps starting from $\mathcal{M}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Following Ho et al. [21], we train the model by minimizing the Mean-Squared Error (MSE) between predicted and added noise:

$$\mathcal{L} = \mathbb{E}_{\mathcal{M}_0 \sim q(\mathcal{M}_0|\mathcal{C}), t\sim[1,T]} ||\epsilon - \epsilon_\theta(\mathcal{M}_t, t, \mathcal{C})||_2^2. \quad (2)$$

The conditional terms $\mathcal{C} = \{\mathcal{M}_{\text{prev}}, \mathcal{M}_{\text{ori}}, \mathcal{E}, \mathcal{P}\}$ comprise: (i) two frames of motion $\mathcal{M}_{\text{prev}}$ right before the current segment, encoded via MLP without noise processes; (ii) the original motion segment $\mathcal{M}_{\text{ori}}$; (iii) the editing instruction encoded through CLIP [48]; and (iv) a progress indicator $\mathcal{P}$ representing the normalized starting frame position within the edited motion [30] using sinusoidal positional encoding [59].

To strengthen the model's adherence to editing instructions, we use classifier-free guidance [20] with weight w:

$$\tilde{\epsilon}_\theta(\mathcal{M}_t, t, \mathcal{C}) = (1+w)\epsilon_\theta(\mathcal{M}_t, t, \mathcal{C}) - w\epsilon_\theta(\mathcal{M}_t, t, \mathcal{C}'), \quad (3)$$

where $\mathcal{C}' = \{\mathcal{M}_{\text{prev}}, \mathcal{M}_{\text{ori}}, \varnothing, \mathcal{P}\}$ represents the conditional terms with the instruction removed.

5.2. Body Part Coordinator

Training on composed motion data introduces a critical challenge: generated motions may exhibit incorrect coordination patterns, such as synchronized movement of same-side feet and hands during walking. To address this, we introduce a motion discriminator D that provides classifier guidance to the diffusion model, ensuring natural coordination between body parts.

The discriminator is trained to classify motion segments as either coherent (uncomposited) or artificially composed. We construct a training dataset where 50% of samples come from unmodified source motion segments in the HumanML3D dataset [18], while the remaining 50% are synthetically created by compositing body parts from different motion segments. Through this balanced training approach, the discriminator learns to identify subtle coordination patterns that distinguish natural from composited motions.

During the motion generation process, we integrate the trained discriminator as a classifier guidance:

$$\tilde{\mathcal{M}}_0 = \hat{\mathcal{M}}_0 + \lambda \nabla_{\hat{\mathcal{M}}_0} D(\hat{\mathcal{M}}_0), \tag{4}$$

where $\hat{\mathcal{M}}_0 = \tilde{\epsilon}_\theta(\mathcal{M}_t, t, \mathcal{C})$ is the model's output, $\tilde{\mathcal{M}}_0$ represents the motion segment after applying classifier guidance, λ controls the guidance strength, and $\nabla_{\hat{\mathcal{M}}_0} D(\hat{\mathcal{M}}_0)$ is the discriminator's gradient with respect to $\hat{\mathcal{M}}_0$. To refine body part coordination while preserving the overall motion structure, we apply this classifier guidance during the final 20 steps of the auto-regressive sampling process.

6. Experiments

6.1. Evaluation Settings

Tasks and Datasets Our main experiments evaluate two key tasks: body part replacement (semantic edits) and style transfer (style edits), as detailed in Sec. 4.1. We assess all methods using our task-specific datasets, split into training (80%), validation (5%), and testing (15%) sets. For training data preparation, we create triplets (original motion, edited motion, instruction) from our STANCE dataset using composition rules in Sec. 4.3. The training set of HumanML3D [18] serves as our extensive motion base for MotionCutMix implementations. The evaluation of fine-grained adjustment capabilities is presented separately in Appendix C.5.

Additionally, we evaluate our method on the MotionFix dataset [7]. For these experiments, we disable MotionCutMix and configure our auto-regressive diffusion model with a 16-frame window size.

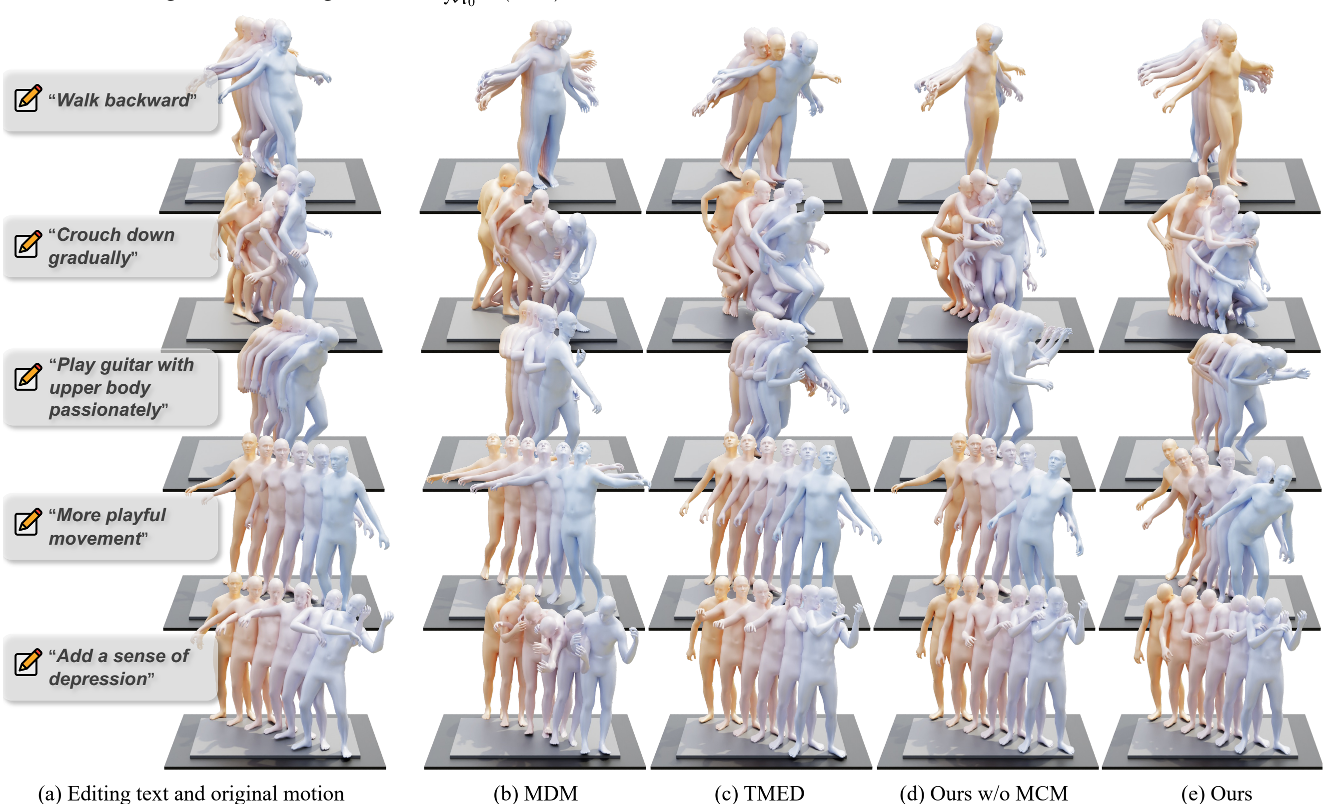


Figure 5. **Qualitative comparison of text-guided motion editing results.** Each sequence shows the original motion alongside edits by MotionReFit and baseline methods. Motion trajectories are visualized with a color gradient from **orange** (starting position) to **blue** (ending position), with spatial offsets applied to emphasize motion differences.

Table 1. **Quantitative comparison across body part replacement (upper) and style transfer (lower) tasks.** Each metric reports mean over 10 evaluations with 95% confidence intervals (±). Arrows (→) indicate metrics where values closer to real data are better. **Bold** denotes best performance.

Method	FID↓	Diversity→	FS↓	Edited-to-Source Retrieval				Edited-to-Target Retrieval			
				R@1→	R@2→	R@3→	AvgR→	R@1↑	R@2↑	R@3↑	AvgR↓
Real Data	$0.01^{\pm.001}$	$36.06^{\pm.436}$	$0.98^{\pm.000}$	$52.08^{\pm.371}$	$54.32^{\pm.314}$	$56.00^{\pm.365}$	$8.28^{\pm.045}$	$100.0^{\pm.000}$	$100.0^{\pm.000}$	$100.0^{\pm.000}$	$1.00^{\pm.000}$
MDM-BP [56]	$0.44^{\pm.030}$	$36.71^{\pm.701}$	$0.91^{\pm.003}$	$69.11^{\pm.912}$	$79.75^{\pm.711}$	$85.14^{\pm.561}$	$2.20^{\pm.028}$	$39.05^{\pm.469}$	$46.39^{\pm.441}$	$50.57^{\pm.556}$	$8.92^{\pm.033}$
TMED [7]	$0.52^{\pm.034}$	$35.37^{\pm.540}$	$0.90^{\pm.008}$	$38.59^{\pm1.169}$	$44.10^{\pm.932}$	$48.67^{\pm.911}$	$9.31^{\pm.211}$	$42.70^{\pm1.533}$	$52.89^{\pm1.286}$	$58.32^{\pm1.430}$	$6.47^{\pm.118}$
TMED w/ MCM	$0.54^{\pm.028}$	$35.67^{\pm.482}$	$0.90^{\pm.006}$	$41.29^{\pm.631}$	$46.13^{\pm.881}$	$49.80^{\pm.945}$	$9.38^{\pm.095}$	$50.62^{\pm1.612}$	$61.95^{\pm1.421}$	$68.52^{\pm1.484}$	$\mathbf{4.48^{\pm.119}}$
Ours w/o MCM	$0.23^{\pm.026}$	$36.34^{\pm.620}$	$0.96^{\pm.003}$	$93.17^{\pm.273}$	$96.30^{\pm.178}$	$97.33^{\pm.206}$	$1.27^{\pm.011}$	$51.18^{\pm.206}$	$53.71^{\pm.275}$	$55.30^{\pm.371}$	$8.51^{\pm.020}$
Ours w/o BC	$0.23^{\pm.016}$	$36.18^{\pm.523}$	$\mathbf{0.97^{\pm.003}}$	$52.51^{\pm.595}$	$\mathbf{56.03^{\pm.368}}$	$58.19^{\pm.358}$	$\mathbf{7.54^{\pm.038}}$	$60.78^{\pm.471}$	$67.17^{\pm.457}$	$71.11^{\pm.521}$	$4.74^{\pm.042}$
Ours full	$\mathbf{0.20^{\pm.025}}$	$\mathbf{36.01^{\pm.758}}$	$\mathbf{0.97^{\pm.002}}$	$\mathbf{52.48^{\pm.337}}$	$56.13^{\pm.361}$	$\mathbf{58.59^{\pm.234}}$	$7.46^{\pm.034}$	$\mathbf{61.37^{\pm.457}}$	$\mathbf{68.35^{\pm.493}}$	$\mathbf{72.20^{\pm.314}}$	$4.65^{\pm.029}$
Real Data	$0.01^{\pm.001}$	$33.98^{\pm.865}$	$0.98^{\pm.000}$	$50.94^{\pm1.791}$	$62.88^{\pm.925}$	$67.40^{\pm.828}$	$6.28^{\pm.058}$	$100.0^{\pm.000}$	$100.0^{\pm.000}$	$100.00^{\pm.000}$	$1.00^{\pm.000}$
MDM-BP [56]	$0.39^{\pm.033}$	$\mathbf{33.64^{\pm.835}}$	$0.89^{\pm.010}$	$62.40^{\pm1.977}$	$82.78^{\pm1.100}$	$89.62^{\pm1.156}$	$1.96^{\pm.062}$	$38.89^{\pm2.152}$	$53.51^{\pm1.167}$	$60.24^{\pm1.122}$	$7.14^{\pm.071}$
TMED [7]	$1.54^{\pm.093}$	$34.37^{\pm1.111}$	$0.90^{\pm.010}$	$28.44^{\pm1.156}$	$40.03^{\pm1.173}$	$46.53^{\pm1.280}$	$8.48^{\pm.104}$	$24.76^{\pm1.440}$	$38.33^{\pm2.067}$	$45.62^{\pm.934}$	$8.12^{\pm.099}$
TMED w/ MCM	$0.84^{\pm.060}$	$34.35^{\pm.669}$	$0.92^{\pm.004}$	$39.83^{\pm1.522}$	$55.00^{\pm1.608}$	$\mathbf{62.92^{\pm1.463}}$	$5.37^{\pm.112}$	$33.02^{\pm1.024}$	$47.60^{\pm1.303}$	$56.94^{\pm1.242}$	$6.15^{\pm.072}$
Ours w/o MCM	$0.23^{\pm.017}$	$34.05^{\pm1.077}$	$0.93^{\pm.006}$	$87.05^{\pm1.345}$	$98.33^{\pm.556}$	$99.41^{\pm.313}$	$1.16^{\pm.012}$	$51.39^{\pm1.406}$	$63.58^{\pm1.058}$	$67.88^{\pm.699}$	$7.15^{\pm.102}$
Ours w/o BC	$0.16^{\pm.018}$	$34.51^{\pm.681}$	$\mathbf{0.95^{\pm.003}}$	$45.52^{\pm1.146}$	$57.05^{\pm1.120}$	$62.29^{\pm.810}$	$6.57^{\pm.080}$	$62.26^{\pm1.838}$	$74.69^{\pm.814}$	$79.90^{\pm1.227}$	$3.51^{\pm.081}$
Ours full	$\mathbf{0.14^{\pm.015}}$	$34.19^{\pm.865}$	$0.94^{\pm.004}$	$\mathbf{47.67^{\pm1.099}}$	$\mathbf{57.71^{\pm1.039}}$	$62.50^{\pm.439}$	$\mathbf{6.46^{\pm.086}}$	$\mathbf{63.82^{\pm1.551}}$	$\mathbf{76.35^{\pm.988}}$	$\mathbf{80.69^{\pm1.009}}$	$\mathbf{3.48^{\pm.062}}$

Baeslines We compare our method against two text-guided motion editing baselines: MDM-BP [56] and TMED [7]. MDM-BP extends the original MDM by incorporating body-part inpainting and ground-truth body part information to specify fixed and edited parts. For TMED comparisons, we maintain their original experimental settings (detailed in Appendix C.2).

Ablations We conduct the following ablation studies to analyze key components of our method:

- Ours w/o MCM: To isolate the impact of motion composition, we evaluate our method using fixed original-edited pairs following SINC [6], without MotionCutMix during training.
- TMED [7] w/ MCM: To assess MotionCutMix's broader applicability, we integrate it into TMED's [7] training pipeline.
- Ours w/o BC: To validate our body part coordinator, we evaluate our method without the body part coordinator from Sec. 5.2.
- MotionCutMix Ratio: To examine data composition effects, we vary the proportions of motion base data used in MotionCutMix.
- Annotated Data Size: To evaluate MotionCutMix with limited annotations, we train models with different proportions of annotated data.
- Window Size: To optimize temporal processing, we experiment with different sliding window sizes for auto-regressive generation.
- Training steps: To assess data randomness effects on convergence, we track performance under different MotionCutMix ratios during training.

Metrics We employ the Edited-to-Source Retrieval (E2S) and Edited-to-Target Retrieval (E2T) scores from Athanasiou et al. [7], using TMR [43] features. We report R@1, R@2, R@3, and AvgR with 32-batch random gallery sampling from the test set. For quality and diversity assessment, we use Fréchet Inception Distance (FID), Foot Score (FS) [71], Diversity, and Multimodality [56]. E2S interpretation varies by task: high scores are desired for fine-grained adjustments and MotionFix evaluations, while body part replacement and style transfer should match reference dataset distributions (detailed in Appendix C.3).

6.2. Comparison Results

Quantitative results in Tab. 1 demonstrate that our full method achieves superior performance across most metrics for both style and semantic edits. The retrieval scores indicate precise editing while preserving the original context. Qualitative results in Fig. 5 showcase our approach's versatile editing capabilities. In semantic editing, our method successfully executes backward walking and crouching while maintaining upper body movements, whereas baseline methods fail to produce coherent motions. The style transfer examples highlight our method's sophisticated control, achieving pronounced style modifications while preserving the original motion's semantic content.

Tab. 2 presents batch-wise evaluation results on the MotionFix benchmark. Even without MotionCutMix augmentation, our auto-regressive approach outperforms TMED and MDM-BP across all metrics. By processing long sequences through fixed-length windows, our method

Table 2. **Quantitative comparison with TMED [7] evaluated on MotionFix dataset [7] using a gallery size of 32.** Results show means across 10 evaluation runs, with **bold** indicating best result.

Method	Edited-to-Source Retrieval				Edited-to-Target Retrieval			
	R@1↑	R@2↑	R@3↑	AvgR↓	R@1↑	R@2↑	R@3↑	AvgR↓
Real Data	74.01	84.52	89.91	2.03	100.0	100.0	100.0	1.00
MDM-BP [56]	61.28	69.55	73.99	4.21	39.10	50.09	54.84	6.46
TMED [7]	71.77	84.07	89.52	1.96	62.90	76.51	83.06	2.71
Ours w/o MCM	**83.47**	**90.42**	**92.84**	**1.73**	**66.33**	**80.05**	**84.98**	**2.64**

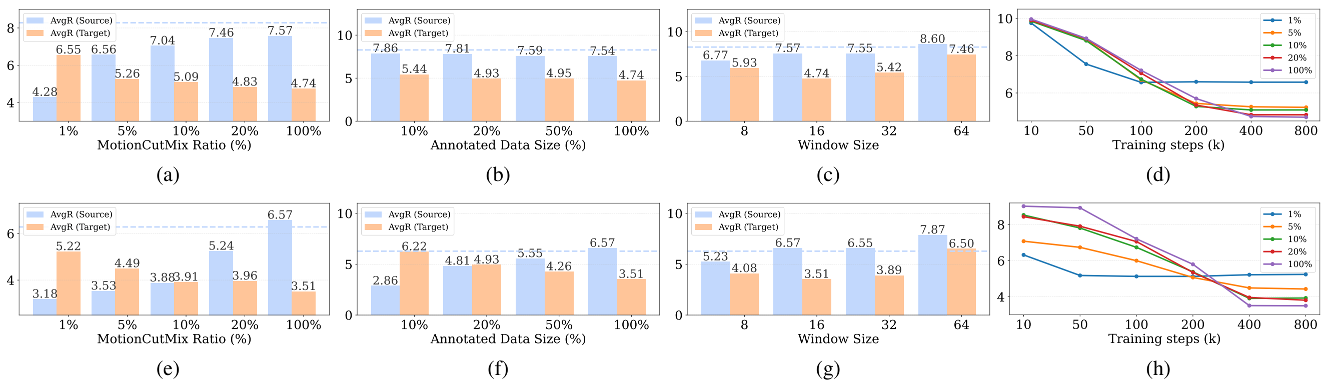
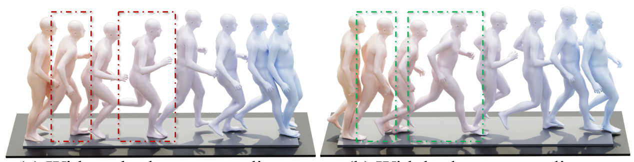


Figure 6. **Ablation analyses for body part replacement (a-d) and style transfer (e-h), reporting AvgR metrics.** Edited-to-Target AvgR shown only for (d) and (h), with blue dotted lines indicating real data Edited-to-Source AvgR. Parameters studied: (a,e) MotionCutMix ratio, (b,f) annotated data volume, (c,g) temporal window size, and (d,h) convergence patterns at varying MotionCutMix ratios. All training converges within 800k steps.



(a) Without body part coordinator (b) With body part coordinator

Figure 7. **Impact of body part coordinator on motion quality.** Examples show paired results using identical random seeds, highlighting how coordinator prevents unnatural synchronous movements of same-side limbs (arm and leg moving forward together).

achieves both higher E2T scores for accurate editing and better E2R scores for context preservation. This demonstrates the effectiveness of our auto-regressive architecture over single-step approaches. Complete evaluation results at full-test set scale are provided in Appendix C.6.

6.3. Ablation Results

Our experiments show that MotionCutMix significantly improves performance for both our method and TMED (Tab. 1), demonstrating its broad applicability to motion editing tasks. The quality improvements from our body part coordinator are visible in Fig. 7 and quantitatively supported by improved FID scores. Importantly, we find that merely learning from composed data is insufficient for proper body part coordination. Quantitative evaluation of guidance strength λ and guidance steps count is presented in Appendix C.4.

Our analysis through Figs. 6a and 6e reveals that performance directly scales with the amount of augmented data—increasing the MotionCutMix Ratio leads to substantial gains in motion editing capabilities. When examining data efficiency in Figs. 6b and 6f, we find that models with MotionCutMix maintain strong performance even with reduced data scales compared to baseline models, indicating reduced dependence on annotated data volume. For temporal processing, our experiments in Figs. 6c and 6g identify 16 frames as the optimal window size, effectively balancing data randomness with motion coherence. Training dynamics shown in Figs. 6d and 6h demonstrate that despite introducing random variations, higher MotionCutMix ratios consistently improve performance without compromising training convergence.

7. Conclusion

This work introduces MotionReFit, a text-guided motion editing framework that enables precise modification of body parts and temporal segments while maintaining motion authenticity. We enhance the framework with MotionCutMix for dynamic training augmentation and incorporate a body part coordinator for movement synchronization. Additionally, we contribute STANCE, a new MoCap and re-annotated dataset targeting three fundamental editing tasks: body part replacement, fine-grained adjustment, and style transfer.

Our work shows that for a specific motion editing task, minimal annotated data is sufficient. Moreover, by reducing the need for high-quality data (*e.g.* MoCap data), our approach opens up broader applications. Specifically, we demonstrate that MotionReFit extends beyond motion editing to interactive modifications and complex compositional motion generation in Appendix E.

Limitations Our approach exhibits limitations in processing long-term temporal dependencies and lacks spatial awareness for position-dependent instructions. A comprehensive discussion on limitations and future directions is provided in Appendix F.

Acknowledgment This work is supported in part by the National Natural Science Foundation of China (62376031), the Beijing Nova Program, the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62, 2020. 2

[2] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64, 2020. 2

[3] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[4] Kenji Amaya, Armin Bruderlin, and Tom Calvert. Emotion from motion. In *Graphics interface*, 1996. 2

[5] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *International Conference on 3D Vision (3DV)*, 2022. 2

[6] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 7, A4

[7] Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *ACM SIGGRAPH Conference Proceedings*, 2024. 1, 3, 4, 6, 7, A2, A3

[8] Xiaoyu Bie, Wen Guo, Simon Leglaive, Lauren Girin, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Hit-dvae: Human motion generation via hierarchical transformer dynamical vae. *arXiv preprint arXiv:2204.01565*, 2022. 2

[9] Matthew Brand and Aaron Hertzmann. Style machines. In *ACM SIGGRAPH Conference Proceedings*, 2000. 2

[10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[11] Ziyi Chang, Edmund J. C. Findlay, Haozheng Zhang, and Hubert P. H. Shum. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. In *Proceedings of International Conference on Computer Graphics Theory and Applications*, 2022. 2

[12] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, A5

[13] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[14] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[15] Yuzhu Dong, Andreas Aristidou, Ariel Shamir, Moshe Mahler, and Eakta Jain. Adult2child: Motion style transfer using cyclegans. In *ACM SIGGRAPH Conference Proceedings*, 2020. 2

[16] Michael Gleicher. Motion path editing. In *Proceedings of Symposium on Interactive 3D Graphics*, 2001. 2

[17] Purvi Goel, Kuan-Chieh Wang, C Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. In *ACM SIGGRAPH Conference Proceedings*, 2024. 3

[18] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 6, A4

[19] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60, 2020. 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5, A1

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5, A2

[22] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):138, 2016. 2

[23] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):161, 2022. 2

[24] Eugene Hsu, Kari Pulli, and Jovan Popović. Style translation for human motion. *ACM Transactions on Graphics (TOG)*, 24(3):1082, 2005. 2

[25] Yue Huang, Haoran Mo, Xiao Liang, and Chengying Gao. Unpaired motion style transfer with motion-oriented projection flow network. In *International Conference on Multimedia and Expo (ICME)*, 2022. 2

[26] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision (ECCV)*, 2025. 1, 3

[27] Leslie Ikemoto, Okan Arikan, and David Forsyth. Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics (TOG)*, 28(1):1, 2009. 2

[28] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3):33, 2022. 2

[29] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[30] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[31] Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Wei Yang, and Li Yuan. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic

graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[32] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2

[33] Manmyung Kim, Kyunglyul Hyun, Jongmin Kim, and Jehee Lee. Synchronized multi-character motion editing. *ACM Transactions on Graphics (TOG)*, 28(3):79, 2009. 2

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. A1

[35] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018. 2

[36] Noah Lockwood and Karan Singh. Biomechanically-inspired motion path editing. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2011. 2

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. A3

[38] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. A2

[39] Wanli Ma, Shihong Xia, Jessica K Hodgins, Xiao Yang, Chunpeng Li, and Zhaoqi Wang. Modeling style and variation in human motion. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2010. 2

[40] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, A1

[42] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[43] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 4, 7, A3

[44] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[45] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[46] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[47] Ziyun Qian, Zeyu Xiao, Zhenyi Wu, Dingkang Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, Dongliang Kou, and Lihua Zhang. Smcd: High realism motion style transfer via mamba-based diffusion. *arXiv preprint arXiv:2405.02844*, 2024. 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 5

[49] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[50] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[51] Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with auto-regressive motion diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):143, 2024. 2

[52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015. 2

[53] Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. Lgtm: Local-to-global text-driven human motion diffusion model. In *ACM SIGGRAPH Conference Proceedings*, 2024. 2

[54] Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. Style-erd: Responsive and coherent online motion style transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[55] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[56] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 7, A2, A3

[57] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[58] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. In *ACM SIGGRAPH Conference Proceedings*, 1995. 2

[59] A Vaswani. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5, A1

[60] Jack M Wang, David J Fleet, and Aaron Hertzmann. Multi-factor gaussian process models for style-content separation. In *International Conference on Machine Learning (ICML)*, 2007. 2

[61] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse

stochastic human-action generators by learning smooth latent transitions. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[62] Andrew Witkin and Zoran Popovic. Motion warping. In *ACM SIGGRAPH Conference Proceedings*, 1995. 2

[63] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34 (4):119, 2015. 2

[64] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[65] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[66] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[67] Wenjie Yin, Yi Yu, Hang Yin, Danica Kragic, and Mårten Björkman. Scalable motion style transfer with constrained diffusion generation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 2

[68] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 4

[69] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(6):4115–4128, 2024. 2

[70] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 3

[71] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International Conference on Computer Vision (ICCV)*, 2023. 7

[72] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *International Conference on Computer Vision (ICCV)*, 2023. 2