This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Cross-Axis Feature Fusion with Joint-Wise Motion Difference Prediction for Text-Based 3D Human Motion Editing

Gyojin Han
School of Electrical Engineering, KAIST
hangj0820@kaist.ac.kr

Junmo Kim
School of Electrical Engineering, KAIST
junmo.kim@kaist.ac.kr

Abstract

We address text-based 3D human motion editing, where the goal is to preserve the style and structure of a source motion while applying edits described in natural language. The release of the MotionFix dataset has spurred active research into training-based diffusion models that directly generate an edited motion from a source motion and a text instruction. While previous works have focused primarily on learning when an edit should occur temporally, our goal is to create a model that understands not only this temporal aspect but also which specific joints are responsible for the change. Targeting this, we propose a novel architecture and a complementary auxiliary task to aid its training. Our architecture consists of two axis-anchored transformers, which extract distinct features along the joint and time dimensions respectively, and a cross-axis fusion block that integrates these representations. We further introduce an auxiliary task that trains the joint-anchored transformer to regress the Soft-DTW distance between source and target joint rotations. This objective teaches the module to understand which joints to modify and which to preserve. Through comprehensive experiments on the MotionFix dataset, we demonstrate that our method significantly improves semantic alignment with both the text instruction and the source motion, as well as the overall fidelity of the generated motion, achieving state-of-the-art results.

1. Introduction

3D human motion generation [4, 5, 8, 12, 23, 30, 32, 41, 44, 45, 48, 50, 51] and editing [3, 4, 11, 22, 26] play central roles in content creation for animation, film, and games. While generation enables rapid prototyping of new sequences from textual prompts, editing is particularly practical when teams must adapt an existing take without discarding its strengths. For example, one may wish to preserve an animator's style or a captured actor's trajectory while making targeted adjustments late in production, such as changing which arm lifts, shifting the timing of a turn, or slightly exaggerating a jump. In these scenarios, text-based editing offers a concise interface for specifying the desired modifications while maintaining continuity with the source sequence.

Early approaches [1, 6, 9, 10, 19, 25, 36, 42] to motion editing often relied on task-specific modules for stylization, pose editing, or in-betweening and non-textual inputs like spatial or temporal constraints. While these methods improved controllability, they lacked the flexibility of text-based instructions and offered limited joint-level specificity. The advent of diffusion models [15, 16, 21, 29, 31, 40] introduced powerful text-to-motion generation capabilities, yet most systems [4, 5, 8, 23, 30, 41, 45, 48, 50, 51] focused on synthesizing motion from scratch rather than minimally modifying an existing sequence.

A significant shift occurred with the release of the MotionFix dataset [3]. By providing triplets of text instruction, source motion, and target motion, it enabled a supervised setting for the editing task. The accompanying baseline, TMED [3], operates in this setting by concatenating a CLIP-based text embedding [37] with a shallow projection of the source motion, feeding the resulting tokens directly to a diffusion transformer as a conditioning signal. Building on this, SimMotionEdit [26] proposed a condition transformer to fuse text and motion features before they are passed to the diffusion transformer (DiT) [31]. SimMotionEdit also introduced an auxiliary motion similarity prediction task to help the model learn when in the sequence changes should occur. However, the SimMotionEdit architecture aggregates features along the joint dimension for each time step, which constrains the extraction of disentangled joint-level information. Consequently, its auxiliary supervision is also restricted to the frame level (predicting frame-wise similarity). While this provides temporal guidance, it offers limited insight into which specific joints are responsible for the change. Therefore, the encoder's understanding of joint-level control remains under-specified, limiting the semantic alignment between the condition and the resulting edits.

To address this limitation, we introduce a new archi-

tecture composed of two axis-anchored transformers and a cross-axis fusion block, together with a complementary joint-wise auxiliary task. Our architecture is designed to separately model the two primary axes of motion data. First, a joint-anchored transformer aggregates the global trajectory for each joint across the entire sequence. Concurrently, a time-anchored transformer captures the full-body pose characteristics for each individual frame. These two distinct representations are then integrated by a cross-axis fusion block, which allows frame-wise features to attend to the joint-wise context. The resulting fused features are used to condition the diffusion transformer for generation. To further guide the joint-anchored module, our joint-wise auxiliary task regresses the per-joint Soft-DTW distance between the source and target rotation trajectories, using the features from the joint-anchored transformer. This auxiliary objective thereby trains the joint-anchored module to comprehend which specific joints must be modified and which must be preserved to accurately reflect the text instruction. We validate the approach through extensive experiments and ablations on the MotionFix dataset, demonstrating consistent improvements over prior methods.

We can summarize the contributions of this work as follows:

- We propose a novel conditioning architecture featuring axis-anchored transformers to separately model the spatial and temporal axes of motion. These are integrated by a cross-axis fusion block to provide joint-aware and temporally-aware conditioning for the diffusion model.
- We introduce a joint-wise motion difference prediction objective, an auxiliary task that uses Soft-DTW over rotation trajectories. This provides timing-robust, joint-specific supervision, guiding the model to learn which joints to modify and how much they should change.
- We conduct comprehensive validation on the MotionFix dataset, including extensive ablation studies that analyze the efficacy of our proposed architecture, supervision design, and fusion strategy.

2. Related Works

2.1. 3D Human Motion Generation

Early approaches to 3D human motion generation relied on recurrent models, variational autoencoders, and generative adversarial networks [12, 32, 44]. With the emergence of diffusion models and rapid gains in their generative capability, diffusion-based text-to-motion methods [4, 5, 8, 23, 30, 41, 45, 48, 50, 51] trained on large-scale paired text–motion datasets [13, 34, 35] have become the dominant paradigm. This shift has delivered steady improvements in fidelity and diversity across benchmarks. Modern pipelines typically predict sequences of parametric human body poses (e.g., SMPL parameters) and often incorporate auxiliary controls such as trajectories, foot contacts, or end-effector constraints [27, 43, 45]. Recent advances further extend conditioning beyond text to shape- or scene-aware signals [46, 47], and explore online or interactive control regimes for real-time applications [51]. Collectively, these works establish a progression from autoregressive and latent-variable generators to diffusion-driven models that integrate multimodal conditioning and explicit control signals for scalable, high-quality motion synthesis.

2.2. 3D Human Motion Editing

Building on progress in generation, research on human motion editing has advanced rapidly, aiming to transform an input motion according to user intent while preserving identity- or style-consistent attributes [3, 4, 11, 22, 26]. Earlier editing methods typically imposed spatial or temporal constraints [9, 10, 25, 42], or decomposed authoring into stylization, pose editing, and in-betweening subproblems [1, 6, 19, 36]; these designs improved controllability but required non-text inputs and offered limited joint-level specificity. Subsequent diffusion-based editing introduced masked inpainting [24], attention manipulation [4], and keyframe-guided heuristics [11], which enhanced directability yet continued to rely on handcrafted interfaces. Text-driven editing then emerged either at the pose level [6] or through indirect prompting via large language models [20, 49], but these approaches still struggled to realize fine-grained, joint-aware edits across full-body sequences. The MotionFix dataset [3] addresses this gap by providing instruction–source–target triplets that enable direct supervision for text-based editing, catalyzing conditional diffusion methods that leverage both the source motion and textual instructions. Within this trajectory, TMED [3] operationalizes conditional diffusion on MotionFix to align edits with natural-language instructions, while SimMotionEdit [26] integrates motion-similarity prediction and a new transformer design to strengthen content preservation and instruction faithfulness during editing. Going beyond prior methods, our approach explicitly specifies which joints should change and how by introducing joint-anchored and time-anchored transformers with a cross-axis fusion block and a joint-wise Soft-DTW auxiliary objective, enabling instruction-faithful, joint-aware edits.

2.3. Diffusion Models

Diffusion models [16] were first popularized in image synthesis, where iterative denoising from Gaussian noise—together with improved parameterizations, samplers, and guidance—yielded large gains in fidelity and controllability [7, 38]. Key developments include score-based/denoising diffusion formulations, accelerated samplers, classifier-free guidance, and transformer-based backbones that scale well with data and compute [15, 29, 31, 40].

The same principles were subsequently extended beyond images to video and 3D content, demonstrating that diffusion in an appropriate latent or parameter space can model complex spatio–temporal structure [14, 17, 18, 39].

3. Method

In this section, we present the architecture and training methodology of our proposed text-based 3D human motion editing model. The model consists of a joint-anchored transformer, a time-anchored transformer, and a cross-axis fusion block for encoding features from the source motion and the text instruction. For the motion generation process, we employ a DiT. For the effective training of this architecture, we introduce a novel auxiliary task, termed joint-wise motion distance prediction. An overview of our method is illustrated in Fig. 1, and a detailed elaboration of each component will be provided in the following subsections.

3.1. Diffusion Model for Human Motion Editing

The recent release of the MotionFix dataset [3], which provides triplets of text instructions, source motions, and target motions, has facilitated recent progress in text-based 3D human motion editing. This dataset allows direct training of diffusion models for this task, and our work focuses on this setting. In this setting, the source motion $S = \{s^i\}_{i=1}^N$ and target motion $T = \{t^i\}_{i=1}^N$ are represented as sequences of human body poses, where N denotes the number of frames in each sequence. Following a previous work [3, 26], we represent each source pose $s^i \in \mathbb{R}^K$ and target pose $t^i \in \mathbb{R}^K$ as a vector with $K = 207$ dimensions. Specifically, this 207-dimensional vector comprises: 126 (21×6) 6D rotation parameters for 21 joints of the SMPL model (excluding the left hand, right hand, and root joint); 66 (22×3) position parameters for 22 joints (all SMPL joints excluding the left and right hands); 3 global translation parameters; and 12 global orientation parameters.

Our goal is to generate an edited motion $M = \{m^i\}_{i=1}^N$ from an input source motion S and a text prompt c using a diffusion model. For this, the diffusion model g is trained to minimize the noise-prediction loss:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\tau,\epsilon}\left[\left\|g\left(T_\tau; f(S,c), \tau\right) - \epsilon\right\|_2^2\right], \tag{1}$$

where τ is a diffusion timestep sampled uniformly from $\{1, \ldots, \mathcal{T}\}$, $\epsilon \sim \mathcal{N}(0, I)$, and f is the feature encoder for the source motion S and text prompt c. More specifically, the noisy target motion T_τ is constructed by the forward diffusion process $T_\tau = \sqrt{\bar{\alpha}_\tau} T + \sqrt{1-\bar{\alpha}_\tau}\epsilon$. where $\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s$ and $\alpha_s = 1 - \beta_s$, with a variance schedule $\{\beta_s\}_{s=1}^{\mathcal{T}}$. At inference time, we synthesize the edited motion by running the DDPM reverse process from pure noise. We initialize $M_{\mathcal{T}} \sim \mathcal{N}(0, I)$ and, for $\tau = \mathcal{T}, \ldots, 2$, we compute

$$\hat{\epsilon}_\tau = g\left(M_\tau; f(S,c), \tau\right), \tag{2}$$

and sample

$$M_{\tau-1} = \frac{1}{\sqrt{\alpha_\tau}}\left(M_\tau - \frac{1-\alpha_\tau}{\sqrt{1-\bar{\alpha}_\tau}}\hat{\epsilon}_\tau\right) + \sqrt{\tilde{\beta}_\tau} z_\tau, \tag{3}$$

where $z_\tau \sim \mathcal{N}(0, I)$ and $\tilde{\beta}_\tau = \dfrac{1-\bar{\alpha}_{\tau-1}}{1-\bar{\alpha}_\tau}\beta_\tau$. After iterating to $\tau = 1$, we take the final edited motion as $M = M_0$.

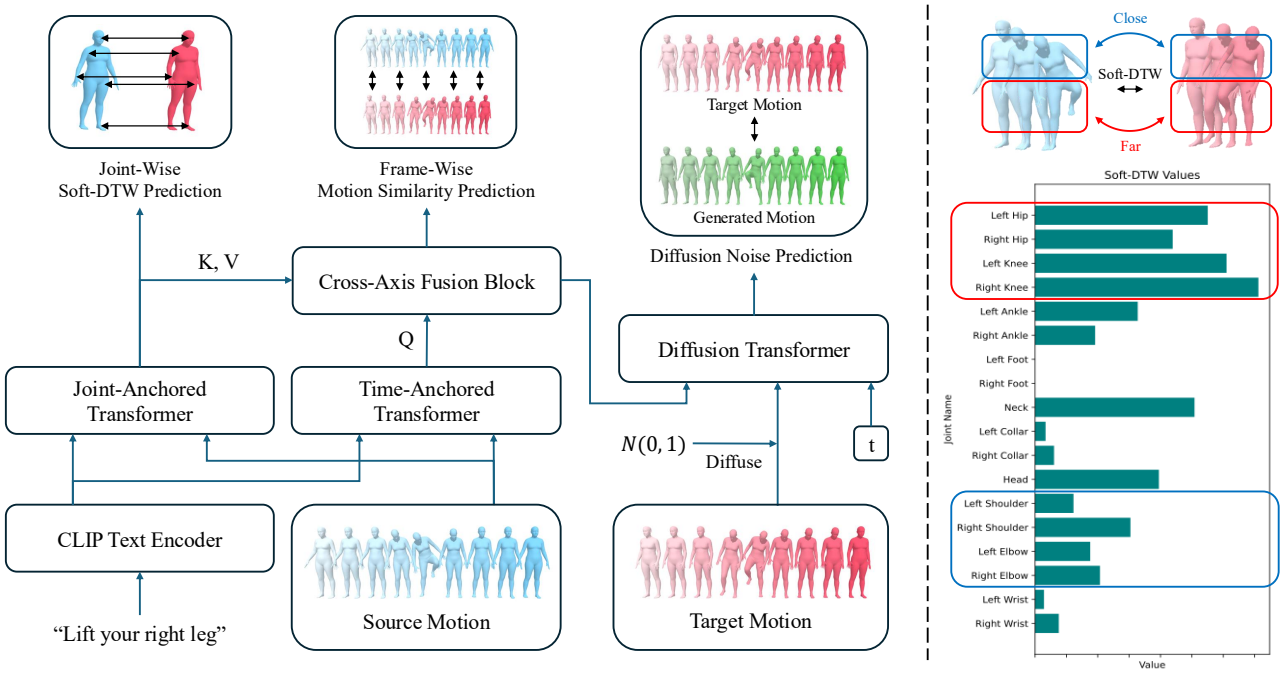
3.2. Axis-Anchored Transformers and Cross-Axis Fusion Block

The generated motion M should preserve the structural and stylistic properties of the source motion S while integrating the edit specified by the text instruction c. To this end, the feature encoder f must (1) establish fine-grained cross-modal alignment between S and c, (2) disentangle edit-relevant factors from content that should remain unchanged, and (3) deliver the resulting conditioning signals to the diffusion model so that denoising applies the intended edit without drifting from S. We now examine the design of f, the conditioning pathway, and the associated training objectives in detail.

In TMED [3], the feature encoder, f_{tmed}, processes the text instruction using a CLIP text encoder and the source motion using a single linear projection layer. The resulting output embeddings are then concatenated. Notably, no explicit fusion occurs between the text and source motion embeddings within the encoder itself. Instead, information exchange is deferred until these embeddings are fed as tokens into the DiT, where fusion occurs via self-attention mechanisms between the text embedding, the source motion embedding, and the noised motion.

SimMotionEdit [26] introduced a condition transformer as a feature encoder f_{sim}, to process both text and source motion embeddings, attempting feature fusion prior to the DiT. Specifically, SimMotionEdit projects the K dimensions of each source pose into a latent dimension D. It then fuses these projected features with the text embedding to obtain an enhanced text feature $h_{\text{text}} \in \mathbb{R}^{1,D}$ and an enhanced motion feature $h_{\text{motion}} \in \mathbb{R}^{N,D}$, where pose characteristics are aggregated per frame. Furthermore, SimMotionEdit trained its condition encoder with an auxiliary task: predicting the frame-wise motion similarity between the source and target motions. This was designed to help the model understand when (at which temporal step) the motion modification should occur. By feeding these enhanced h_{text} and h_{motion} as conditioning signals into the DiT, SimMotionEdit achieved results that demonstrated better alignment with the text instruction compared to TMED.

However, while the SimMotionEdit approach may help in understanding when in the sequence a change should occur, its reliance on predicting a global, frame-wise similarity metric provides limited information regarding where in the body the edit should be applied. The model is trained



(a) Overview of the proposed method

(b) Soft-DTW values

Figure 1. Overview of the proposed approach and joint-wise supervision. **(a)** A joint-anchored transformer and a time-anchored transformer produce enhanced features, h_{joint} and h_{time}. A cross-axis fusion block performs multi-head attention with these and the fused representation conditions a DiT that predicts denoising updates to generate the edited motion from noise, given the source motion and the text. Two auxiliary heads supervise the encoders by frame-wise motion similarity and joint-wise Soft-DTW prediction. **(b)** Example Soft-DTW values between the source and target rotation trajectories for some joints. We observe that joints exhibiting larger motion changes yield higher Soft-DTW distances than those with smaller changes.

to predict the overall similarity difference per frame, which does not help it learn which specific joints should undergo significant modification versus which joints should remain consistent with the source motion. For instance, if the source motion at frame n involves lifting the left leg, and the text instruction is "Lift your right leg," SimMotionEdit's architecture might help identify that a change is needed at frame n. However, it provides no explicit guidance for the model to understand that the hip and knee joints require substantial modification, while the upper body joints should remain relatively unchanged.

We claim that for the text-based motion editing task, capturing the global characteristics of each joint across the entire sequence is as important as understanding the temporal characteristics of the pose at each frame. Therefore, we attempt to represent these two axes distinctly using two anchored transformers. Both the joint-anchored transformer and the time-anchored transformer take the source motion $S \in \mathbb{R}^{N \times K}$ and the text instruction c as input. They output a joint-wise feature $h_{\text{joint}} \in \mathbb{R}^{K,D}$ and a frame-wise feature $h_{\text{time}} \in \mathbb{R}^{N,D}$, respectively. Following this, we introduce a cross-axis fusion block—implemented as a multi-head attention block—to integrate the information from both features. This block receives h_{time} as the Query and h_{joint} as both the Key and Value. The resulting fused feature, h_{fusion}, is then concatenated with the noised motion input to the DiT, thereby conditioning the diffusion process on both the source motion and the text instruction. Through this architecture comprising the joint-anchored transformer, the time-anchored transformer, and the cross-axis fusion block, we have designed a feature encoder f_{fusion} that comprehends not only when the motion should be altered but also which specific joints require modification. The architecture of our model is illustrated in Fig. 1 (a).

3.3. Joint-Wise Motion Difference Prediction

In addition to the above architecture, we introduce an auxiliary task that encourages the joint-anchored transformer to learn joint-specific dynamics as intended. This task regresses, for each rotation channel, a scalar measure of the motion difference between the source motion S and the target motion T. Because joint motion is best represented by rotation parameters, we first extract from the joint-wise feature h_{joint} the sub-tensor corresponding to the 21 SMPL joints with 6D rotations, denoted $h'_{\text{joint}} \in \mathbb{R}^{K' \times D}$, where $K' = 126$. The regression head φ_{reg} takes h'_{joint} and out-

puts a per-channel score that should reflect how much the rotation trajectory changes from S to T while being insensitive to timing shifts. Because h'_{joint} aggregates information over time and many text edits change the start time or speed without altering motion morphology, a frame-by-frame loss would over-penalize harmless timing shifts. We therefore need a distance that is robust to local time warps and compares motions by trajectory shape rather than absolute timing. To provide such supervision, we use Soft-Dynamic Time Warping (Soft-DTW) as the ground-truth target.

Specifically, let $S', T' \in \mathbb{R}^{K' \times N}$ be the rotation-only versions of S and T. For each rotation parameter $j \in \{1, \dots, K'\}$, define the time series $S'_j = (s_{j,1}, \dots, s_{j,N})$ and $T'_j = (t_{j,1}, \dots, t_{j,N})$. The regression head predicts a scalar score per channel,

$$\hat{d}_j \;=\; \big[\varphi_{\text{reg}}(h'_{\text{joint}})\big]_j \in \mathbb{R}, \qquad j = 1, \dots, K'. \tag{4}$$

We construct the ground-truth score using dynamic time warping. Classical DTW compares two sequences up to local time shifts by minimizing the accumulated pointwise cost over monotone alignment paths. With $d(x_n, y_m) = \|x_n - y_m\|_2^2$,

$$\text{DTW}(x, y) \;=\; \min_{\pi \in \mathcal{A}} \sum_{(n,m) \in \pi} d(x_n, y_m), \tag{5}$$

where $\mathcal{A}$ is the set of admissible monotone warping paths. Soft-DTW replaces the hard minimum with a smooth softmin (log-sum-exp) controlled by temperature $\gamma > 0$,

$$\text{SoftDTW}_\gamma(x, y) \;=\; \text{softmin}^{(\gamma)}_{\pi \in \mathcal{A}} \sum_{(n,m) \in \pi} d(x_n, y_m), \tag{6}$$

which preserves DTW's invariance to local time warps while being fully differentiable. We adopt Soft-DTW in practice and use available GPU-accelerated implementations for efficient training.

Our auxiliary supervision targets the Soft-DTW distance between the source and target rotation trajectories for each channel,

$$d_j \;=\; \text{SoftDTW}_\gamma\big(S'_j,\, T'_j\big) \in \mathbb{R}, \tag{7}$$

and trains the regression head to match these values from h'_{joint} via a mean-squared loss,

$$\mathcal{L}_{\text{aux}} \;=\; \frac{1}{K'} \sum_{j=1}^{K'} \big(\hat{d}_j - d_j\big)^2. \tag{8}$$

This objective guides the joint-anchored transformer to encode joint-wise motion differences that capture similarity in trajectory shape while remaining robust to differences in start time and duration, complementing the main diffusion objective. As an example, we visualize the Soft-DTW measurements for a sample as a graph in Fig. 1 (b).

4. Experiments

4.1. Experimental Setting

Dataset and baselines. We utilize the MotionFix dataset [3] for the training and evaluation of our model. This dataset, which is designed for text-based 3D motion editing, provides a total of 6,730 annotated samples. Each data point is structured as a triplet, containing a source motion, a target motion, and a corresponding text instruction. The dataset was constructed by first leveraging the TMR motion embedding space [33] to select semantically similar motion pairs from motion capture datasets. Following this, human annotators manually generated concise descriptions detailing the differences between the paired motions.

We designate TMED [3], MotionReFit [22] and SimMotionEdit [26] as our direct baselines, which are state-of-the-art (SOTA) methods trained on the same MotionFix dataset [3] capable of generating an edited motion solely from free-form text and a source motion. Additionally, following previous works, we also compare our proposed method against MDM [41] and MDM-BP [3]. MDM is a text-to-motion generation model, and MDM-BP is an adaptation of MDM for the text-based motion editing task that extracts body part labels via GPT [2].

Evaluation metrics. To evaluate our method, we follow the motion-to-motion retrieval-based benchmark proposed in MotionFix [3], which utilizes the pre-trained TMR [33] as a feature extractor. This benchmark assesses how well the edited motions are semantically aligned with both the text instructions and the source motions by checking how well the corresponding target motions and source motions are retrieved, respectively. Specifically, we compute the top-1, top-2, and top-3 retrieval accuracies (R@1, R@2, R@3), along with the average rank, within a batch size of 32. Furthermore, following the official implementation of SimMotionEdit, we also report these metrics on the test set for a more comprehensive evaluation. To assess the fidelity and diversity of the generated results, we additionally measure the Fréchet Inception Distance (FID) between our generated motions and the ground-truth target motions.

Implementation details. We extract text embeddings from the text instructions for editing using a pre-trained CLIP [37] (ViT-B/32) as previous works [3, 26]. The joint-anchored transformer and the time-anchored transformer share an identical structure of 4 transformer encoder layers, 8 attention heads, and a 512-dimensional latent space. The diffusion transformer is deeper, built with 8 transformer encoder layers, 8 attention heads, and a 512 latent dimension. To merge the feature representations from the axis-anchored transformer modules, we designed a cross-axis fusion block, which consists of a multi-head cross-attention layer also configured with 8 attention heads and a 512 latent dimension. For training, we utilized the AdamW op-

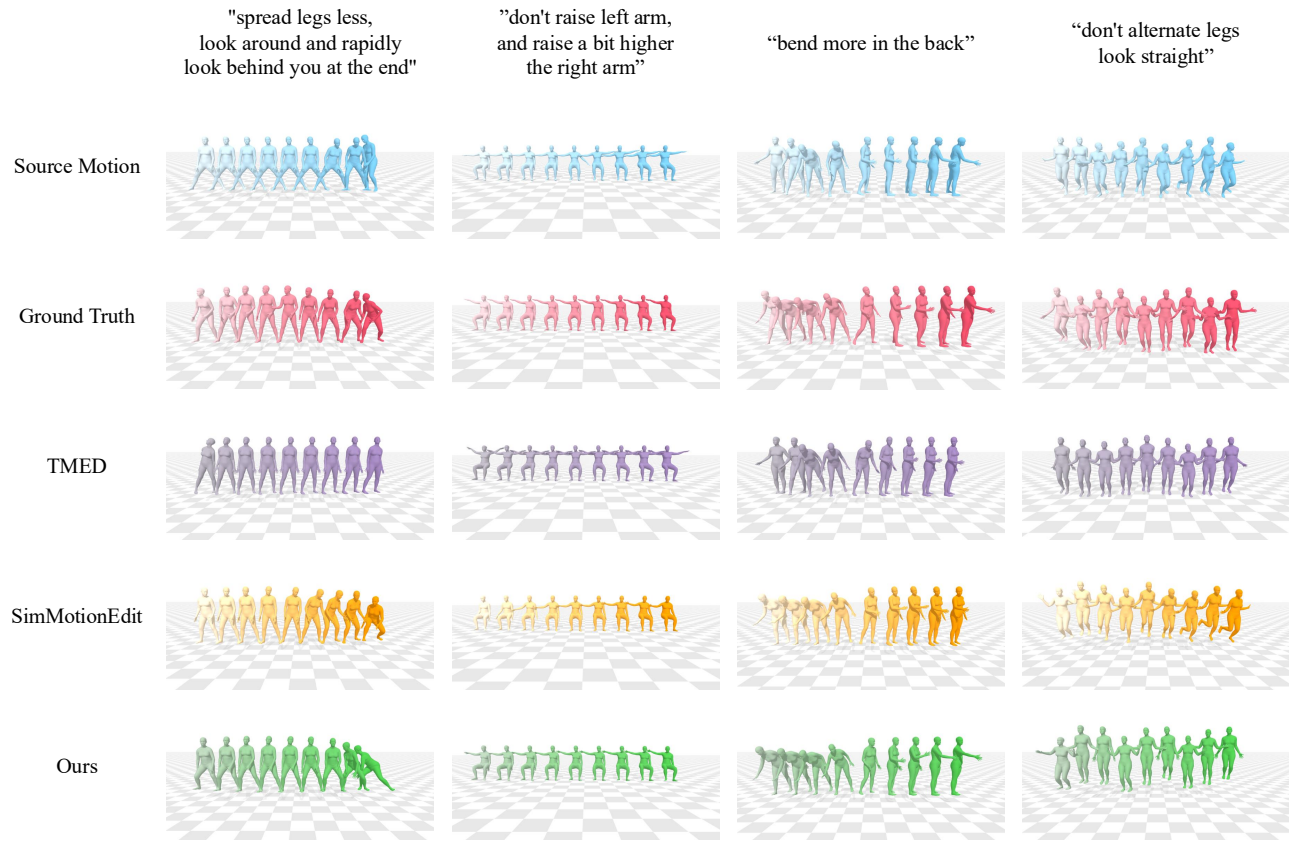


Figure 2. **Qualitative results.** We visualize the source motion, ground truth, and the edited motions from our method and competing methods, given a text instruction. To effectively illustrate the temporal progression, rendered meshes are translated to the right over time. For each motion, frame recency is encoded by saturation: lower saturation represents earlier frames, while higher saturation indicates more recent frames. Best viewed zoomed in.

timizer [28] with a learning rate of 10^{-4} and a batch size of 64. The model was trained for 1,500 epochs, a process that required approximately 12 hours on a single NVIDIA RTX 4090 GPU. During the sampling phase, we adopted the inference configuration from prior studies [3, 26]. We employed the DDPM [16] sampling process fixed at 300 diffusion steps, paired with a cosine noise scheduler. A two-way conditioning strategy was applied, setting the guidance scale to 2.0 for both the text and source motion inputs.

4.2. Experimental Results

Quantitative results. As shown in Table 1, the quantitative results demonstrate that the proposed method achieves state-of-the-art performance, outperforming all baselines across all retrieval-related metrics on both the batch and test set evaluations. Specifically, our method achieves the highest Top-K retrieval accuracies (R@1, R@2, and R@3). The superior R@1 score is particularly significant, as it demonstrates that the motion edited by our model is most frequently the single closest match to the ground truth target motion. This indicates a high degree of precision in adhering to the editing instructions. Furthermore, our method obtains the best (lowest) average rank. This result signifies that, on average, the rank of the correct target motion is significantly closer to the top, highlighting the consistent accuracy and robustness of our model's generation capabilities compared to previous works.

In addition to semantic alignment with the text instructions and source motions, the fidelity and diversity of the generated motion are critical for overall quality. We evaluate this using the FID between the edited motions and the target motions. Our model achieves the lowest FID score and this result confirms that the motions generated by our approach are not only semantically accurate but also exhibit a distribution that is statistically closer to that of real motions, indicating superior realism and variety.

Qualitative results. Recognizing that numerical metrics alone cannot fully capture the semantic coherence and perceptual quality of edited motions, we provide a qualitative comparison in Figure 2. This figure visualizes the source motion, ground truth, and the results from our method and competing baselines for various text instructions.

Methods	Generated-to-Target (Batch)				Generated-to-Target (Test Set)				FID↓
	R@1↑	R@2↑	R@3↑	AvgR↓	R@1↑	R@2↑	R@3↑	AvgR↓	
Ground Truth	100.0	100.0	100.0	1.00	64.36	88.75	95.56	1.74	–
MDM	4.03	7.56	10.48	15.55	0.10	0.10	0.10	–	–
MDM-BP	39.10	50.09	54.84	6.46	8.69	14.71	18.36	180.99	–
TMED	62.90	76.51	83.06	2.71	14.51	21.72	28.73	56.63	0.167
MotionReFit	66.33	80.05	84.98	2.64	-	-	-	-	-
SimMotionEdit	70.62	82.92	88.12	2.38	25.49	39.33	49.21	23.49	0.110
Ours	**74.38**	**88.54**	**92.08**	**1.92**	**29.45**	**45.26**	**54.55**	**16.42**	**0.097**

Table 1. **Quantitative results.** We compare the generated-to-target motion-to-motion retrieval performance of the proposed method with that of state-of-the-art text-based motion editing models. We report the top-1 (R@1), top-2 (R@2), and top-3 (R@3) retrieval accuracies along with the average rank (AvgR). To evaluate the fidelity and diversity of the generated motion, we additionally compute the FID between the edited motions and the target motions. Metrics are marked with ↑ (higher is better) or ↓ (lower is better). We highlight the **best** results in bold and the second-best results with an underline.

Motion Sim.	Joint Delta	Generated-to-Target (Batch)				Generated-to-Target (Test Set)				FID↓
		R@1↑	R@2↑	R@3↑	AvgR↓	R@1↑	R@2↑	R@3↑	AvgR↓	
✗	✗	72.08	85.42	89.17	2.13	**30.24**	43.28	51.78	18.58	0.122
✗	L2	71.46	84.38	89.38	2.16	29.25	42.49	52.77	18.80	0.113
✗	Soft-DTW	72.92	86.25	90.00	2.07	28.85	41.70	50.59	18.26	0.108
✓	✗	71.04	85.42	90.21	2.03	29.45	44.07	54.35	17.98	0.143
✓	L2	71.04	86.88	90.21	1.97	26.48	39.72	49.01	17.54	0.113
✓	Soft-DTW	**74.38**	**88.54**	**92.08**	**1.92**	29.45	**45.26**	**54.55**	**16.42**	**0.097**

Table 2. We conducted an ablation study to analyze the impact of auxiliary tasks on the text-based human motion editing performance. We evaluated the contributions of both the motion similarity prediction (Motion Sim.) task and the joint-wise distance prediction task (Joint Delta) by experimenting with whether each task was performed. For the joint-wise motion distance prediction task, we additionally conducted an experiment using the L2 distance as the distance metric, comparing it against Soft-DTW. Metrics are marked with ↑ (higher is better) or ↓ (lower is better). We highlight the **best** results in bold and the second-best results with an underline.

As shown in the figure, TMED often struggles to accurately reflect the given text instructions, resulting in motions that are not semantically aligned with the edit. While both SimMotionEdit and our proposed method demonstrate significant improvements over TMED, our approach consistently generates motions that are more faithful to the textual commands.

Furthermore, our method achieves stronger alignment with the source motion, retaining its original characteristics while precisely implementing the desired changes. We attribute this advantage to our model's architecture. We introduce an auxiliary task that trains the joint-anchored transformer to explicitly predict how each joint's motion should be altered (i.e., how much and in what way) according to the text instruction. The features from this transformer are then fused and provided as conditional input to the DiT. This mechanism enables our model to better discern which elements of the source motion must be preserved and which parts require editing.

4.3. Ablation Study

Ablation study on the auxiliary tasks. We conduct an ablation study to validate the effectiveness of our proposed auxiliary task and to analyze its individual impacts on performance. For the joint-wise motion distance prediction task, we investigate three variants: (1) not using the task (✗), (2) using L2 distance as the distance metric, and (3) using Soft-DTW as the distance metric on the rotation parameters.

As shown in Table 2, employing a simple L2 distance for the joint-wise motion distance prediction yields no significant performance improvement over the baselines that omit this task. We hypothesize that this is because the L2 loss is highly sensitive to temporal variations. For instance, if a motion has the similar semantic shape but is performed slightly faster or slower, the L2 distance can fluctuate dramatically. This volatility prevents the joint-anchored transformer from effectively learning to distinguish between temporal variations and genuine semantic

γ for Soft-DTW	Generated-to-Target (Batch)				Generated-to-Target (Test Set)				FID↓
	R@1↑	R@2↑	R@3↑	AvgR↓	R@1↑	R@2↑	R@3↑	AvgR↓	
0.1	73.54	86.46	90.21	2.08	27.87	<u>42.69</u>	51.19	20.13	0.107
0.5	72.08	85.00	<u>91.67</u>	<u>1.93</u>	25.49	41.90	50.40	<u>17.20</u>	0.112
1	73.54	86.25	89.38	2.00	<u>28.06</u>	42.29	<u>52.57</u>	17.81	0.115
5	73.75	<u>87.08</u>	90.83	2.00	26.28	41.90	51.98	18.39	<u>0.106</u>
10	<u>74.38</u>	**88.54**	**92.08**	**1.92**	**29.45**	**45.26**	**54.55**	**16.42**	**0.097**
50	**74.79**	85.21	89.79	2.08	27.87	<u>42.69</u>	51.78	19.21	0.108

Table 3. We analyze the model's performance by varying the γ value for Soft-DTW from 0.1 to 50. Metrics are marked with ↑ (higher is better) or ↓ (lower is better). We highlight the **best** results in bold and the <u>second-best</u> results with an underline.

Condition Feat.	Generated-to-Target (Batch)				Generated-to-Target (Test Set)				FID↓
	R@1↑	R@2↑	R@3↑	AvgR↓	R@1↑	R@2↑	R@3↑	AvgR↓	
Time-anchored	70.62	82.92	87.92	2.28	24.11	37.35	46.05	22.93	0.111
Joint-anchored	<u>72.08</u>	84.58	<u>91.04</u>	2.08	<u>27.67</u>	40.91	<u>50.99</u>	18.82	0.110
Mean	70.42	83.54	89.58	2.15	26.48	40.71	48.62	20.79	0.114
Axis-fusion	70.00	<u>84.79</u>	90.00	<u>2.02</u>	26.09	<u>41.50</u>	48.81	<u>18.54</u>	<u>0.108</u>
Ours	**74.38**	**88.54**	**92.08**	**1.92**	**29.45**	**45.26**	**54.55**	**16.42**	**0.097**

Table 4. Ablation results on additional explicit text conditioning. Ours uses only the implicit fused feature f_{fusion} without additional explicit text guidance. We highlight the **best** results in bold and the <u>second-best</u> results with an underline.

changes required by the instruction. In contrast, using Soft-DTW leads to substantial performance gains across all retrieval metrics. This demonstrates that Soft-DTW, by being robust to temporal misalignments, provides a much more effective and stable learning signal. We believe this significant improvement arises because the Soft-DTW objective helps the model develop a more precise understanding of which joints to preserve from the source motion and which to modify according to the text.

Ablation study on different γ values. We conducted an ablation study to observe the effect of the γ hyperparameter used in the Soft-DTW calculation. In Soft-DTW, γ serves as a smoothing parameter that controls the differentiability of the alignment operation; a smaller γ approaches the non-differentiable standard DTW, while a larger γ results in a smoother approximation. We evaluated the model's performance by varying γ across six different values: 0.1, 0.5, 1.0, 5.0, 10.0, 50.0. As shown in Table 3, the model's performance is not overly sensitive to the choice of γ, showing robust results across a wide range of values.

Effect of additional explicit text condition. In our framework, the DiT model primarily receives text-related information implicitly through the fused feature f_{fusion}. Following previous work [26], we conducted an ablation study by introducing explicit text features from different modules as auxiliary conditioning variables to investigate whether providing explicit text guidance can further improve the model's performance. Specifically, we evaluated the effects of incorporating features from: (1) the time-anchored transformer, (2) the joint-anchored transformer, (3) the element-wise mean of the time-anchored and joint-anchored features, and (4) the axis-fusion transformer.

As shown in Table 4, the results indicate that providing no explicit condition to the DiT achieves the best performance across all metrics. We hypothesize that this result is due to the information from the text instruction already being sufficiently encoded and infused into f_{fusion}, which serves as the primary input to the DiT. This suggests that an additional condition is redundant for the diffusion process.

5. Conclusion

In this work, we presented a 3D human motion editing framework capable of reasoning about the specific joints that drive intended motion updates. Our architecture employs axis-anchored transformers and a cross-axis fusion block to effectively decouple and integrate spatial and temporal cues from both source motion and text instructions. Furthermore, we introduced a joint-wise motion difference objective using Soft-DTW, which allows the model to discern which joints to modify versus preserve while maintaining robustness to temporal shifts. Extensive evaluations on the MotionFix dataset demonstrate significant improvements in semantic alignment and motion fidelity, validating the efficacy of our proposed architecture and auxiliary task.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2024-00439020, Developing Sustainable, Real-Time Generative AI for Multimodal Interaction, SW Starlab).

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Trans. Graph.*, 39(4), 2020. 1, 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[3] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J. Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 3, 5, 6

[4] Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms. *arXiv e-prints*, pages arXiv–2410, 2024. 1, 2

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2023. 1, 2

[6] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: Correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15018–15028, 2023. 1, 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 2

[8] Xuehao Gao, Yang Yang, Zhenyu Xie, Shaoyi Du, Zhongqian Sun, and Yang Wu. Guess: Gradually enriching synthesis for text-driven human motion generation. *IEEE Transactions on Visualization and Computer Graphics*, 30 (12):7518–7530, 2024. 1, 2

[9] Michael Gleicher. Motion editing with spacetime constraints. In *Proceedings of the 1997 Symposium on Interactive 3D Graphics*, page 139–ff., New York, NY, USA, 1997. Association for Computing Machinery. 1, 2

[10] Michael Gleicher. Motion path editing. In *Proceedings of the 2001 Symposium on Interactive 3D Graphics*, page 195–202, New York, NY, USA, 2001. Association for Computing Machinery. 1, 2

[11] Purvi Goel, Kuan-Chieh Wang, C. Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2

[12] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2021–2029, New York, NY, USA, 2020. Association for Computing Machinery. 1, 2

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 2

[14] Gyojin Han, Jiwan Hur, Jaehyun Choi, and Junmo Kim. Learning neural deformation representation for 4d dynamic shape generation. In *Computer Vision – ECCV 2024*, pages 186–203, Cham, 2025. Springer Nature Switzerland. 3

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 1, 2, 6

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646. Curran Associates, Inc., 2022. 3

[19] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4), 2016. 1, 2

[20] 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 *Computer Vision – ECCV 2024*, pages 180–196, Cham, 2025. Springer Nature Switzerland. 2

[21] Jiwan Hur, Jaehyun Choi, Gyojin Han, Dong-Jae Lee, and Junmo Kim. Expanding expressiveness of diffusion models with limited data via self-distillation based fine-tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5028–5037, 2024. 1

[22] Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22735–22745, 2025. 1, 2, 5

[23] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Runyi Yu, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Local action-guided motion diffusion model for text-to-motion generation. In *Computer Vision – ECCV 2024*, pages 392–409, Cham, 2025. Springer Nature Switzerland. 1, 2

[24] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8255–8263, 2023. 2

[25] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 39–48, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 1, 2

[26] Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. Simmotionedit: Text-based human motion editing with motion similarity prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27827–27837, 2025. 1, 2, 3, 5, 6, 8

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[29] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, pages 5775–5787. Curran Associates, Inc., 2022. 1, 2

[30] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27859–27871, 2025. 1, 2

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 1, 2

[32] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, 2021. 1, 2

[33] Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9488–9497, 2023. 5

[34] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 4(4):236–252, 2016. PMID: 27992262. 2

[35] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 2

[36] Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *ACM Trans. Graph.*, 41(6), 2022. 1, 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 5

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[39] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 2

[41] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 5

[42] Andrew Witkin and Michael Kass. Spacetime constraints. In *Proceedings of the 15th Annual Conference on Computer Graphics and Interactive Techniques*, page 159–168, New York, NY, USA, 1988. Association for Computing Machinery. 1, 2

[43] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[44] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, Wenjun Zeng, and Wei Wu. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2228–2238, 2023. 1, 2

[45] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14928–14940, 2023. 1, 2

[46] Kebing Xue and Hyewon Seo. Shape conditioned human motion generation with diffusion model. *arXiv preprint arXiv:2405.06778*, 2024. 2

[47] Jiaxu Zhang, Shaoli Huang, Zhigang Tu, Xin Chen, Xiaohang Zhan, Gang YU, and Ying Shan. Tapmo: Shape-aware motion generation of skeleton-free characters. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[48] Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17592–17602, 2025. 1, 2

[49] 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 Neu-*

ral Information Processing Systems, pages 13981–13992. Curran Associates, Inc., 2023. 2

[50] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. 1, 2

[51] Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2