Text-to-Any-Skeleton Motion Generation Without Retargeting

Qingyuan Liu[1] Ke Lu[1,2] Kun Dong[1] Jian Xue[1*] Zehai Niu[1] Jinbao Wang[3]
[1]University of Chinese Academy of Sciences [2]Peng Cheng Laboratory [3]Shenzhen University
{liuqingyuan23, dongkun22, niuzehai18}@mails.ucas.ac.cn, {luk, xuejian}@ucas.ac.cn
wangjb@szu.edu.cn

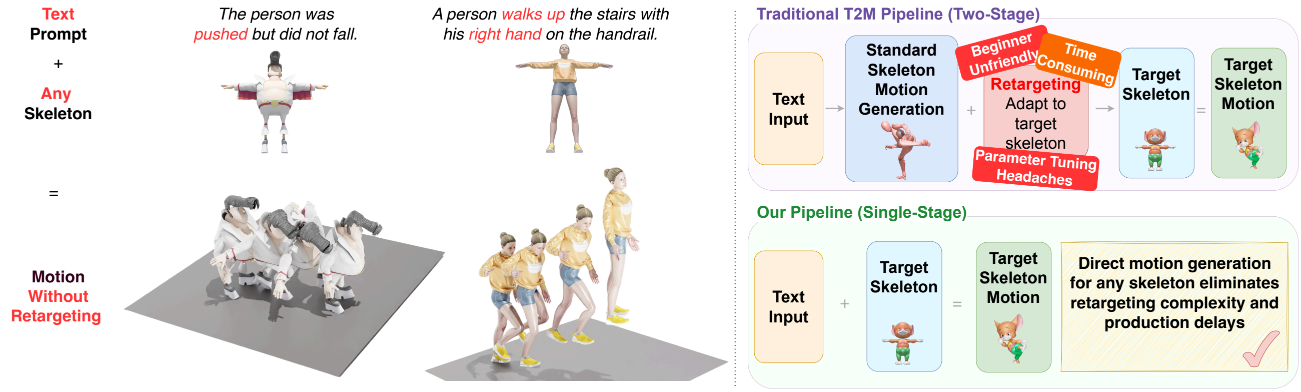
Text Prompt
+
Any Skeleton
=
Motion Without Retargeting
The person was pushed but did not fall.
A person walks up the stairs with his right hand on the handrail.
Traditional T2M Pipeline (Two-Stage)
Text Input
Standard Skeleton Motion Generation
+
Beginner Unfriendly
Time Consuming
Retargeting
Adapt to target skeleton
Parameter Tuning Headaches
Target Skeleton
=
Target Skeleton Motion
Our Pipeline (Single-Stage)
Text Input
+
Target Skeleton
=
Target Skeleton Motion
Direct motion generation for any skeleton eliminates retargeting complexity and production delays

Figure 1. Introducing OmniSkel: A text-driven motion generation framework that **directly generates motions for any skeletons without retargeting**. Unlike traditional two-stage pipelines (top right) that require complex retargeting procedures, our single-stage approach (bottom right) only needs text input and target bone lengths (which are easily obtainable) to generate desired motions. As shown on the left, OmniSkel creates natural movements like "*pushed but did not fall*" and "*walks up stairs with right hand on handrail*" for characters with different body proportions while maintaining action fidelity.

Abstract

Recent advances in text-driven motion generation have shown notable progress. However, these methods are typically limited to standardized skeletons and rely on a cumbersome retargeting process to adapt to varying skeletal configurations of diverse characters. In this paper, we present OmniSkel, a novel framework that directly generates high-quality human motions for any user-defined skeleton without retargeting. Specifically, we introduce a skeleton-aware RVQ-VAE, which utilizes Kinematic Graph Cross Attention (K-GCA) to effectively integrate skeletal information into motion encoding and reconstruction. Moreover, we propose a simple yet effective training-free approach, Motion Restoration Optimizer (MRO), to ensure zero bone length error while preserving motion smoothness. To support this research, we construct SkeleMotion-3D, a large-scale text-skeleton-motion dataset based on HumanML3D. Extensive experiments demonstrate the excellent robustness and generalization of our method.

1. Introduction

Recent years have witnessed remarkable progress in text-to-motion (T2M) generation, showing promising potential in applications such as gaming, metaverse, and virtual/augmented reality [10, 16, 32, 41, 50, 51]. This technology enables the intuitive creation of 3D character animations through natural language descriptions, significantly simplifying content creation workflows in animation [20], virtual reality [16], video games [30, 47], and robotics [5, 22].

Despite these advancements, a fundamental limitation restricts the practical application of current text-to-motion generation methods. As illustrated in Figure 1, existing approaches primarily focus on enhancing the motion quality and text-motion alignment, neglecting variations in skeletons and only generating human motions for standardized skeletons. However, real-world production environments routinely require varying skeletal configurations for diverse

[*] Corresponding author.

characters. To this end, practitioners must resort to skeleton retargeting - a technically demanding process that necessitates significant expertise and frequently fails to produce usable results, inevitably leading to laborious tuning and production delays. Therefore, there is a pressing demand for straightforward motion generation given arbitrary skeletons.

To address this issue, we present OmniSkel, a novel framework that can directly generate high-quality human motion sequences for arbitrary character skeletons. As shown in Figure 1, our approach eliminates the complex retargeting stage required by traditional pipelines, offering a single-stage solution that directly produces character-specific motion without specialized expertise. Specifically, we propose a Skeleton-aware RVQ-VAE (SR-VAE) architecture, which utilizes Kinematic Graph Cross Attention (K-GCA) to effectively integrate skeletal information into the motion encoding and reconstruction. Unlike previous approaches, our design effectively separates skeleton-invariant features, which is crucial for arbitrary skeleton adaptation. To ensure zero bone length error while preserving motion smoothness, we further propose a training-free approach, Motion Restoration Optimizer (MRO). With the above designs, our OmniSkel can deliver high-quality motion generation precisely in line with textual descriptions and character skeletons, providing robust support for practical text-to-motion applications.

To facilitate our research, we construct SkeleMotion-3D, a large-scale dataset based on HumanML3D, pairing textual descriptions with motion sequences across diverse skeletal configurations. This dataset establishes a strong foundation for motion generation with arbitrary skeletons.

Before delving into the details, we summarize our core contributions in this work as follows:

- To the best of our knowledge, we present the first text-to-motion generation framework capable of directly producing motions for arbitrary character skeletons without retargeting, significantly simplifying creation workflows in real-world production environments.
- We propose a novel Skeleton-aware RVQ-VAE (SR-VAE) architecture, which employs Kinematic Graph Cross Attention (K-GCA) to effectively integrate skeletal information into the motion encoding and reconstruction. Moreover, we further introduce Motion Reconstruction Optimizer (MRO) during inference, ensuring zero bone length error while preserving motion smoothness.
- Based on HumanML3D, we construct SkeleMotion-3D, a large-scale text-skeleton-motion dataset. This establishes a foundation for arbitrary skeleton motion generation, facilitating future research in this field.
- To verify the effectiveness of our method, we conduct extensive experiments on challenging datasets. Results demonstrate the superior robustness and generalization of our method. Further analysis reveals the contribution of each component to the performance improvement.

2. Related Work

2.1. Text-to-Motion Generation

Text-to-motion generation has evolved significantly as a vital research area in computer animation and human-computer interaction.

Early Approaches. Initial methods established text-to-motion mappings through alignment techniques [3, 4, 9, 26, 36, 46], but produced motions that lacked naturalism and diversity due to their deterministic nature.

Probabilistic Frameworks. VAE-based approaches [6, 16, 28, 31, 32] model distributions of possible motions, while vision-language alignment methods [27, 33, 37, 40] leveraged pre-trained models for better semantic consistency.

Recent Advances. Contemporary approaches include diffusion-based models [7, 10, 12, 21, 38, 41, 51, 52] generating high-fidelity motions through progressive denoising. Vector Quantized VAE (VQ-VAE) approaches treat motions as discrete token sequences, including autoregressive models [19, 45, 50, 54, 55], bidirectional methods [18, 35], and hybrid approaches [34]. Part-based generation methods [13, 54, 55] provide control over individual body components.

Limitations. Existing approaches remain constrained to standard skeletal structures, creating a gap between research and industry needs where adaptability to various character rigs is essential.

Our work addresses this limitation by developing a skeleton-agnostic framework that generates semantically consistent motions for arbitrary skeletal structures.

2.2. Motion Retargeting

Motion retargeting adapts existing motion data between characters with different proportions and structures. Historically, Gleicher *et al.* [14] pioneered this field through spacetime optimization with kinematic constraints. Following this work, Lee and Shin [23] applied inverse kinematics with B-spline curve fitting, while Choi and Ko [11] simultaneously developed an online method preserving high-frequency details.Tak and Ko [39] later enhanced these approaches by utilizing dynamics constraints for physical plausibility.

With the advent of deep learning, new approaches emerged, including Villegas *et al.* [44] with unsupervised retargeting via recurrent networks, Lim *et al.* [25] disentangling pose and movement, and Aberman *et al.* [1] designing skeleton-aware networks for cross-structural retargeting. More recently, research has shifted towards preserving motion semantics. For instance, Zhang *et al.* [49] developed a residual retargeting network with skeleton-aware

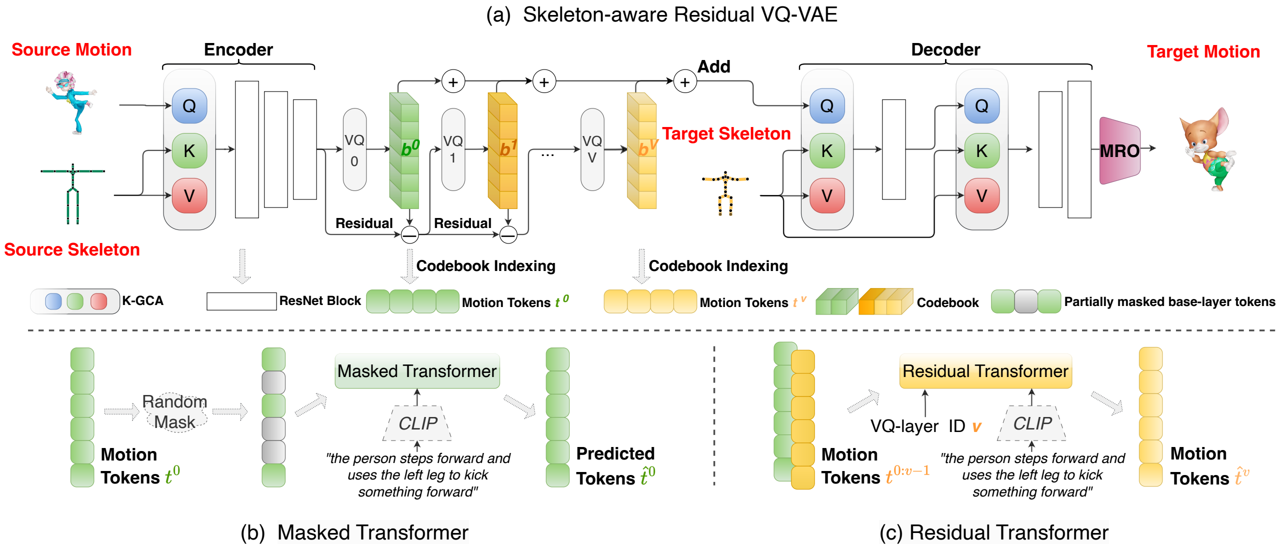

Figure 2. **Method overview.** (a) Skeleton-aware Residual VQ-VAE: our framework processes source motion and skeleton through K-GCA (Kinematic Graph Cross Attention), which divides the input into Q, K, and V branches, creating a hierarchical representation with motion-invariant features. The residual quantization approach progressively captures motion details across multiple layers, enabling high-fidelity reconstruction for arbitrary target skeletons. (b) Masked Transformer: base-layer motion tokens t^0 are randomly masked during training and the text-conditioned transformer learns to predict these tokens simultaneously, capturing fundamental motion patterns. (c) Residual Transformer: builds upon the base layer by progressively predicting higher-level residual tokens t^v conditioned on both text and previous layer tokens, enhancing motion fidelity and skeleton-specific details.

and shape-aware modules. Similarly, Lee *et al*. [24] proposed Skeleton-Agnostic Motion Embedding (SAME) using graph convolutional networks to separate skeleton information from motion while preserving semantics. Building upon these semantic approaches, Zhang *et al*. [48] introduced Semantics-aware Motion reTargeting (SMT), leveraging vision-language models to extract and maintain motion semantics through a two-stage pipeline with skeleton-aware pre-training and semantic constraints.

However, despite this significant progress, many approaches still struggle with maintaining semantic consistency across dramatically different skeletal structures.

3. Approach

Our goal is to generate motion sequences $\mathbf{M}_{1:T}$ of length T conditioned on both a text description c and a target skeletal structure $\mathbf{S}$, where $\mathbf{M}_t \in \mathbb{R}^{J \times D_\text{m}}$ with J denoting the number of joints and D_m representing the dimension of joint features. The skeletal structure $\mathbf{S} \in \mathbb{R}^{J \times D_\text{s}}$ characterizes the static joint properties of the target character. As illustrated in Figure 2, our framework consists of two main components: a skeleton-aware VAE module that enables motion generation for arbitrary skeletal structures (Section 3.2), and a cooperative transformer architecture for motion token prediction (Section 3.3). We first introduce our newly proposed dataset and its representation (Section 3.1) to facilitate better understanding of our method. Then, we detail the complete training pipeline (Section 3.3) and inference process (Section 3.4).

3.1. SkeleMotion-3D Dataset

Our dataset comprises three key components: textual descriptions, motion sequences, and corresponding skeletal structures. Building upon the HumanML3D dataset [16], which combines motion data from AMASS [29] and HumanAct12 [15], we significantly expand the motion-skeleton pairs through a skeleton randomization process inspired by [24].

Feature Representation. We represent motion sequences as $\mathbf{M} \in \mathbb{R}^{T \times J \times D_\text{m}}$, where the motion feature F_motion with dimension D_m consists of:

$$F_\text{motion} = [F_\text{topo}, F_\text{root}, F_\text{xyz}, F_\text{rot}, F_\text{vel}], \tag{1}$$

comprising skeletal topology features, root joint features, relative spatial positions, 6D rotation features, and velocity features.

The skeletal structures are encoded as $\mathbf{S} \in \mathbb{R}^{J \times D_\text{s}}$, where the skeletal feature F_skel with dimension D_s consists of:

$$F_\text{skel} = [F_\text{topo}, F_\text{static}], \tag{2}$$

where F_topo represents topology features and F_static denotes static skeletal properties.

Dataset Construction. Following the preprocessing pipeline established by HumanML3D, we standardize the raw motion sequences through temporal normalization (20 FPS), sequence cropping (maximum 10 seconds), skeletal retargeting, and orientation alignment (Z+ direction). The text annotations from HumanML3D, collected via Amazon Mechanical Turk with quality control measures, are preserved in our dataset.

Dataset Statistics. The SkeleMotion-3D dataset contains 14,616 distinct motion patterns, 29,233 different skeletal configurations, 43,848 unique text-motion-skeleton triplets, and 44,970 textual descriptions with 5,371 unique words. This represents a substantial increase compared to the training dataset (949 motion patterns and 4,745 motion-skeleton pairs) used in [24]'s motion retargeting network. The significant expansion enhances the robustness and generalization capability of our skeleton-aware motion generation model.

3.2. Skeleton-aware Residual VQ-VAE

We propose Skeleton-aware Residual VQ-VAE (SR-VAE), a novel architecture that extends the conventional Residual VQ-VAE to enable skeleton-conditioned motion generation. Conventional motion VQ-VAEs [17, 19, 50, 53] and recent RVQ-VAEs [18, 45] focus on reconstructing source motions with identical skeletal structures. In contrast, as illustrated in Figure 2, SR-VAE adopts an asymmetric VAE architecture, consisting of an encoder, a decoder, a residual quantization module, and a Kinetic Graph Cross Attention (K-GCA) module, enabling motion reconstruction for arbitrary target skeletons.

Architecture Design. Our SR-VAE processes motion generation through three main stages. First, the input motion sequence $\mathbf{M}_{\text{in}} \in \mathbb{R}^{T \times J \times D_{\text{m}}}$ and its corresponding skeleton structure $\mathbf{S}_{\text{in}} \in \mathbb{R}^{J \times D_{\text{s}}}$ are fused via K-GCA before being encoded into latent features. Following MoMask [18], these features undergo residual quantization to produce $V + 1$ ordered code sequences. Finally, the decoder incorporates target skeleton features $\mathbf{S}_{\text{tgt}} \in \mathbb{R}^{J \times D_{\text{s}}}$ through multiple K-GCA layers at different decoding stages to generate the target motion $\mathbf{M}_{\text{tgt}} \in \mathbb{R}^{T \times J \times D_{\text{m}}}$.

Let $\hat{\mathbf{M}}_{\text{tgt}} \in \mathbb{R}^{T \times J \times D_{\text{m}}}$ denote the predicted motion sequence and $\mathbf{M}_{\text{tgt}}$ denote the ground truth target motion. For residual quantization, let $\mathbf{r}^{v} \in \mathbb{R}^{T \times d}$ represent the residual features at layer $v \in \{0, 1, ..., V\}$ and $\mathbf{b}^{v} \in \mathbb{R}^{T \times d}$ denote the corresponding quantized codes, where d is the dimension of the latent representation. Additionally, let $\mathbf{D}_{\text{xyz}} \in \mathbb{R}^{T \times J \times 3}$ and $\hat{\mathbf{D}}_{\text{xyz}} \in \mathbb{R}^{T \times J \times 3}$ represent the ground truth and predicted joint positions in 3D space, respectively. The overall training objective consists of three terms:

$$\mathcal{L}_{\text{total}} = \|\mathbf{M}_{\text{tgt}} - \hat{\mathbf{M}}_{\text{tgt}}\|_1 + \beta \sum_{v=0}^{V} \|\mathbf{r}^{v} - \text{sg}[\mathbf{b}^{v}]\|_2^2 + \lambda \|\mathbf{D}_{\text{xyz}} - \hat{\mathbf{D}}_{\text{xyz}}\|_1 \tag{3}$$

where the first term is motion reconstruction loss, the second term is commitment loss for residual quantization, and the third term specifically focuses on joint positions accuracy. Here, sg$[\cdot]$ denotes the stop-gradient operation, β and λ are weighting factors for the commitment loss and joint position loss, respectively. The loss function is optimized via a straight-through gradient estimator [42]. Following [18, 35, 50], we employ exponential moving average for codebook updates and reset.

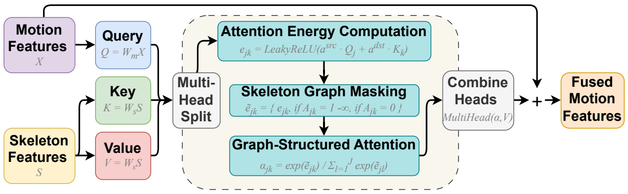
Motion Features
X
Query
$Q = W_m X$
Key
$K = W_s S$
Skeleton Features
S
Value
$V = W_s S$
Multi-Head Split
Attention Energy Computation
$e_{jk} = LeakyReLU(a^{src} \cdot Q_j + a^{dst} \cdot K_k)$
Skeleton Graph Masking
$\tilde{e}_{jk} = \{ e_{jk}, if A_{jk} = 1 -\infty, if A_{jk} = 0 \}$
Graph-Structured Attention
$\alpha_{jk} = exp(\tilde{e}_{jk}) / \Sigma_{l=1}^{J} exp(\tilde{e}_{jl})$
Combine Heads
$MultiHead(\alpha,V)$
+
Fused Motion Features

Figure 3. Architecture of our Kinetic Graph Cross Attention.

Kinetic Graph Cross Attention. To effectively fuse motion and skeletal features, we propose Kinetic Graph Cross Attention (K-GCA), as illustrated in Figure 3, extending the traditional Graph Attention Networks (GAT) [8, 43] for motion-skeleton feature fusion. Unlike standard GAT that performs self-attention within a single feature type, K-GCA enables cross-modal attention between motion and skeletal features while preserving the skeletal graph structure.

Given motion features $\mathbf{X} \in \mathbb{R}^{T \times J \times D_{\text{m}}}$ and skeleton features $\mathbf{S} \in \mathbb{R}^{J \times D_{\text{s}}}$, K-GCA first projects them into a shared space using learnable projection matrices $\mathbf{W}_{\text{m}} \in \mathbb{R}^{D_{\text{m}} \times d_k}$ and $\mathbf{W}_{\text{s}} \in \mathbb{R}^{D_{\text{s}} \times d_k}$, where d_k is the dimension of the projected space:

$$\mathbf{Q} = \mathbf{W}_{\text{m}}\mathbf{X}, \quad \mathbf{K} = \mathbf{V} = \mathbf{W}_{\text{s}}\mathbf{S}. \tag{4}$$

The attention scores between joints j and k are computed using learnable vectors $\mathbf{a}^{\text{src}} \in \mathbb{R}^{d_k}$ and $\mathbf{a}^{\text{dst}} \in \mathbb{R}^{d_k}$:

$$e_{jk} = \text{LeakyReLU}(\mathbf{a}^{\text{src}} \cdot \mathbf{Q}_j + \mathbf{a}^{\text{dst}} \cdot \mathbf{K}_k), \tag{5}$$

To preserve the skeletal structure, we mask the attention scores using the skeleton adjacency matrix $\mathbf{A} \in \{0, 1\}^{J \times J}$, where $\mathbf{A}_{jk} = 1$ if joints j and k are connected in the skeleton graph, and 0 otherwise. We set attention scores for disconnected joints to negative infinity before applying softmax:

$$\tilde{e}_{jk} = \begin{cases} e_{jk}, & \text{if } \mathbf{A}_{jk} = 1 \\ -\infty, & \text{if } \mathbf{A}_{jk} = 0 \end{cases}, \tag{6}$$

$$\alpha_{jk} = \frac{\exp(\tilde{e}_{jk})}{\sum_{l=1}^{J} \exp(\tilde{e}_{jl})}, \tag{7}$$

The final output $\hat{\mathbf{X}} \in \mathbb{R}^{T \times J \times D_{\text{m}}}$ is computed by applying the attention weights to value matrix $\mathbf{V}$ in each head, concatenating results from all heads, and adding back to the input through residual connection:

$$\hat{\mathbf{X}} = \text{LayerNorm}(\mathbf{X} + \text{MultiHead}(\alpha, \mathbf{V})). \tag{8}$$

where $\text{MultiHead}(\alpha, \mathbf{V})$ represents the multi-head weighted combination of values using the computed attention weights α. This design allows K-GCA to capture dynamic relationships between motion and skeletal features while maintaining the anatomical constraints imposed by the skeletal structure. Compared to standard GAT, our approach provides two key advantages: (1) the ability to perform cross-modal attention between motion and skeletal features, and (2) the incorporation of skeletal structure constraints through attention masking.

3.3. Training Pipeline

Our training pipeline follows a sequential approach for skeleton-agnostic motion generation from text, as shown in Figure 2.

SR-VAE Training. First, we train the Skeleton-aware Residual VQ-VAE (SR-VAE) to establish skeleton-agnostic motion representation. This model encodes motion sequences into hierarchical tokens and decodes them back to motion features, independent of skeletal structure. After training, both encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are frozen.

Tokenization of Training Data. Using the frozen SR-VAE encoder, we convert all training motion sequences into hierarchical token representations. For each motion $\mathbf{M}_{\text{in}}$ with corresponding skeleton $\mathbf{S}_{\text{in}}$, we extract multi-level token sequences $\mathbf{t}^{0:V}$ as targets for our text-to-token models.

Masked Transformer Training. Inspired by MoMask [18], we implement a bidirectional Masked Transformer to generate base-layer tokens ($\mathbf{t}^0$). This design offers parallel generation capability and better global motion coherence compared to autoregressive approaches. During training, the model learns to predict original tokens at masked positions from text prompts.

Residual Transformer Training. Finally, we train the Residual Transformer to predict higher-layer tokens ($\mathbf{t}^{1:V}$) encoding motion refinements. Following MoMask [18], this transformer takes both text descriptions and tokens from previous layers as input, focusing on finer motion details while maintaining consistency with the base motion.

3.4. Inference Process

During inference, OmniSkel generates skeleton-aware motions through the following steps:

Base Token Generation: Given a text prompt, the Masked Transformer generates the base-layer tokens $\hat{\mathbf{t}}^0$ that capture primary motion patterns.

Hierarchical Refinement: The Residual Transformer progressively generates higher-layer tokens $\hat{\mathbf{t}}^{1:V}$, starting with layer 1 and proceeding sequentially. Each layer's prediction utilizes both the text prompt and all previously generated token layers.

Motion Decoding: Once we have the complete token hierarchy $\hat{\mathbf{t}}^{0:V}$, we feed these tokens along with the target skeleton structure $\mathbf{S}_{\text{tgt}}$ into the frozen SR-VAE decoder to obtain motion features.

Algorithm 1: Motion Restoration Optimizer

```
Input: Measurement data: $F_{\text{topo}}$, $F_{\text{root}}$, $F_{\text{xyz}}$; Bone
       lengths: $S$
Output: Smooth motion sequence $M_{\text{smooth}}$ with
        guaranteed skeletal consistency
// Step 1: Restore initial motion using
   position-based approach;
$M_{\text{ric}} \leftarrow \text{ric}(F_{\text{topo}}, F_{\text{root}}, F_{\text{xyz}})$;
$M_{\text{smooth}} \leftarrow \varnothing$;
foreach frame $f \in M_{\text{ric}}$ do
    // Step 2: Extract unit direction vectors for
       each bone (except the root);
    Initialize set $\mathcal{U} \leftarrow \varnothing$;
    for $i \leftarrow 1$ to $J - 1$ do
        Let $\mathbf{p}_i$ be the position of joint $i$ in frame $f$;
        Let $\mathbf{p}_{\text{parent}(i)}$ be the position of joint $i$'s
          parent;
        Compute displacement: $\mathbf{v}_i \leftarrow \mathbf{p}_i - \mathbf{p}_{\text{parent}(i)}$;
        Compute unit direction: $\mathbf{u}_i \leftarrow \frac{\mathbf{v}_i}{|\mathbf{v}_i|}$;
        Add $\mathbf{u}_i$ to $\mathcal{U}$;
    end
    // Step 3: Reconstruct the current frame
       using extracted directions, bone lengths, and
       topology;
    $f_{\text{rec}} \leftarrow \text{ReconstructFrame}(\mathcal{U}, S, F_{\text{topo}}, F_{\text{root}})$;
    Add $f_{\text{rec}}$ to $M_{\text{smooth}}$;
end
return $M_{\text{smooth}}$;
```

Feature-to-Motion Conversion: Following the generation of motion features by our SR-VAE decoder, we face the challenge of transforming these abstract representations into production-ready motion data. Contemporary text-to-motion frameworks typically rely on either position-based restoration (offering fluidity but with bone length inconsistencies) or rotation-based restoration (maintaining skeletal integrity but introducing jitter and penetration artifacts).

To overcome these limitations, we employ our **Motion Restoration Optimizer** (MRO, detailed in Algorithm 1), which synergistically combines the strengths of both approaches. MRO guarantees two critical properties simultaneously: absolute preservation of skeletal proportions (zero bone length error) and maintenance of smooth trajectory characteristics, producing physically plausible output $\hat{\mathbf{M}}_{\text{tgt}}$.

Method	FID↓	R Precision↑			MPJPE↓	Skeleton Error↓	Jitter Error↓	Skating Ratio ↓
		Top 1	Top 2	Top 3				
SR-VAE (full model)	**0.027**	**0.508**	**0.700**	**0.793**	0.035	13.778	**1.824**	0.074
w/ Single-head K-GCA	0.040	0.497	0.685	0.789	**0.034**	12.294	2.935	**0.061**
w/o K-GCA	0.065	0.495	0.689	0.785	0.041	17.726	9.594	0.092
w/o RVQ	0.165	0.467	0.639	0.775	0.061	19.726	3.594	0.092
w/ Symmetric Reconstruction	0.885	0.450	0.641	0.745	0.097	89.747	9.867	0.067
w/ Rotation-based Restoration	**0.027**	**0.508**	**0.700**	**0.793**	0.063	**0.00**	30.724	0.193
w/ MRO	**0.027**	**0.508**	**0.700**	**0.793**	0.035	**0.00**	2.084	0.090

Table 1. Ablation study of SR-VAE components tested on the SkeleMotion-3D test set. Our SR-VAE is trained on the SkeleMotion-3D dataset. To evaluate SR-VAE's reconstruction performance for arbitrary skeletal motions, we input random skeleton motions to the encoder and use the standard skeleton from the HumanML3D dataset as the target skeleton.

4. Experiments

We conduct extensive experiments to evaluate our approach across multiple dimensions, focusing on both semantic accuracy and physical plausibility.

Evaluation Metrics. We assess our model using established semantic metrics from prior work [16]: (1) *Frechet Inception Distance* (FID) for overall motion quality; (2) *R-Precision* and *multimodal distance* for text-motion alignment.

To comprehensively evaluate physical plausibility—a critical aspect for animation applications—we introduce several specialized metrics:

- *Skating Ratio*: Measures foot sliding artifacts where feet unnaturally glide across the ground
- *Skeleton Error*: Quantifies bone length variations throughout motion sequences
- *MPJPE* (Mean Per-Joint Position Error): Assesses positional accuracy of joints
- *Jitter Error*: Evaluates motion continuity and natural flow

Detailed formulations of these metrics appear in the supplementary materials.

Implementation Details. Our PyTorch implementation features a Skeleton-aware Residual VQ-VAE (SR-VAE) with residual blocks following T2M-GPT [50] and MoMask [18] architectures. We use a downscale factor of 4, with a residual quantization comprising 6 layers (each containing 512 codes of 512 dimensions) and quantization dropout ratio of 0.2. For the SR-VAE training objective, we set commitment loss weight β to 0.02 and joint position loss weight λ to 2.0. The K-GCA module employs 2-head attention.

Both transformer models consist of 6 layers with 6 attention heads and 384-dimensional latent representations, trained on our SkeleMotion-3D dataset. We use a learning rate of 2e-4 with linear warm-up over 2000 iterations, batch sizes of 512 for SR-VAE and 64 for transformers. During inference, we apply classifier-free guidance (scale 4 for masked transformer, 5 for refinement transformer) with 10 iterative decoding steps. All experiments run on NVIDIA RTX 4090 GPUs.

4.1. Skeleton-Aware Disentanglement Analysis

A cornerstone of our approach is SR-VAE's ability to disentangle motion patterns from skeletal structures. We conduct a targeted experiment to validate this skeleton-agnostic capability.

Figure 4 presents our analysis of four distinct motion patterns, each performed by three different skeletal structures with varying proportions. We process these motion-skeleton pairs through our SR-VAE encoder $\mathcal{E}$ and visualize the resulting base layer tokens $\mathbf{t}^0$ on aligned line charts.

The results reveal that motions with identical semantic patterns but executed by anatomically diverse skeletons produce remarkably similar token sequences. This consistency demonstrates that our encoder $\mathcal{E}$ successfully extracts skeleton-invariant motion features, preserving essential semantic content regardless of the executing skeletal structure.

This confirms that SR-VAE effectively decomposes motion into two orthogonal components: (1) skeleton-invariant pattern information encoded in tokens, and (2) skeleton-specific structural information provided separately to the decoder. This disentanglement enables our model to generate physically plausible animations for arbitrary skeletal structures while preserving intended motion semantics.

4.2. Ablation Study

To validate SR-VAE's skeleton-aware capabilities, we conduct experiments using random input skeletons while decoding to standard HumanML3D [16] skeletons. This evaluates both semantic alignment and cross-skeleton reconstruction capabilities.

Table 1 presents our ablation results on the SkeleMotion-3D test set. The complete SR-VAE model achieves superior performance with an FID of 0.026 and strong R-

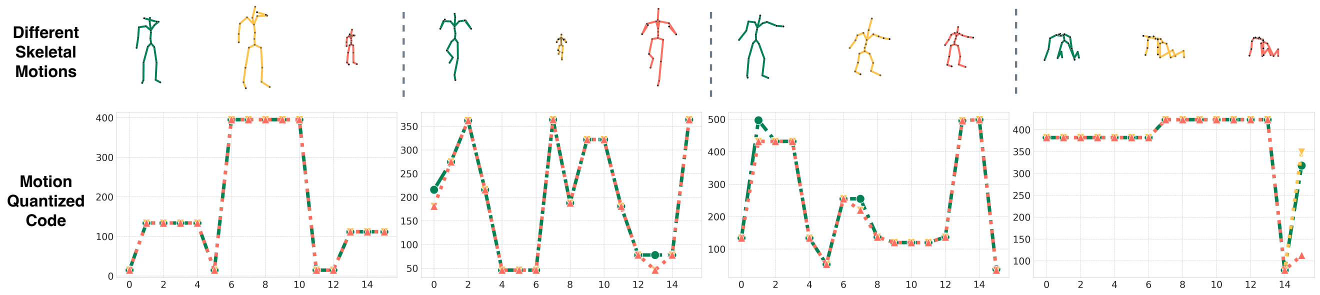

Figure 4. **Motion token analysis across skeletal structures.** This figure demonstrates token sequences generated by the SR-VAE encoder for four distinct motion patterns, each performed by three skeletal structures with different proportions. The consistent token patterns across each motion type (columns) despite skeletal variations demonstrates our encoder's ability to extract skeleton-invariant motion features. Top row shows skeletal visualizations while bottom row displays the corresponding quantized token values, revealing clear pattern preservation across different anatomical configurations.

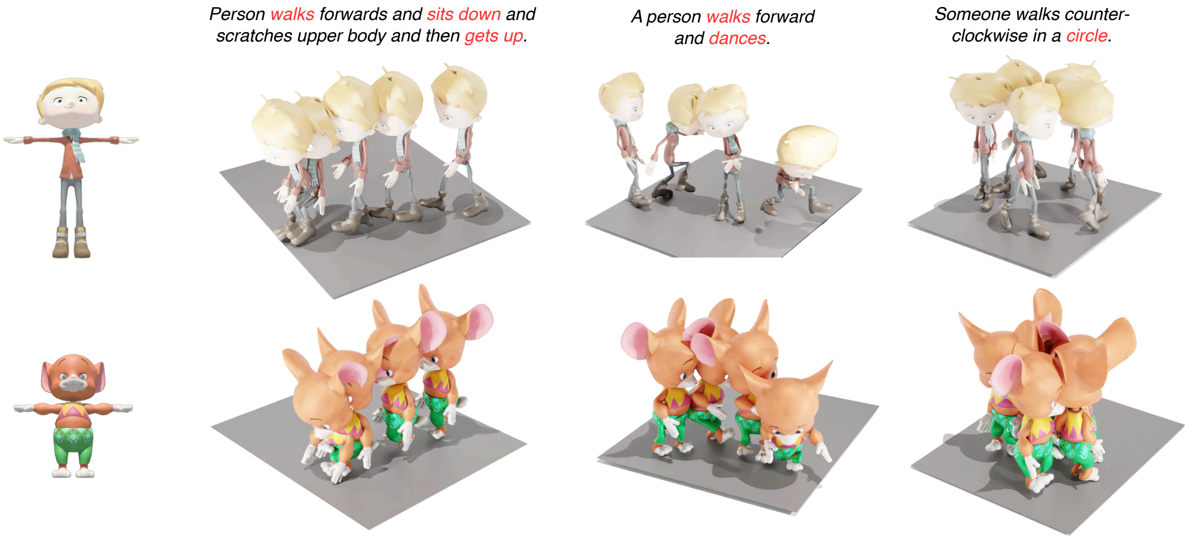
Person walks forwards and sits down and scratches upper body and then gets up.
A person walks forward and dances.
Someone walks counter-clockwise in a circle.

Figure 5. **Text-to-motion generation across different skeletal structures.** Our OmniSkel framework generates diverse motions responding to different text prompts (columns) while adapting to different skeletal anatomies (rows). The human skeleton on the left is randomly selected from Mixamo [2], demonstrating OmniSkel's ability to maintain semantic consistency across dramatically different skeletal structures without retargeting.

Precision scores (0.508/0.700/0.793 for Top-1/2/3), demonstrating excellent semantic consistency.

Architecture Components. Using symmetric reconstruction (same skeleton for encoder and decoder) drastically increases FID to 0.885 and skeleton error to 89.747, confirming that asymmetric processing is essential for effective cross-skeleton learning. Removing K-GCA degrades FID to 0.065 with a significant increase in jitter error (9.594 vs. 1.824), demonstrating its importance in capturing cross-skeleton geometric correspondences.

Without RVQ, FID increases substantially to 0.165 and MPJPE to 0.061, highlighting RVQ's crucial role in maintaining semantic consistency across different skeletal structures. The single-head K-GCA variant shows moderate performance decline (FID: 0.040), indicating that multi-head attention more effectively captures the rich geometric relationships between skeletons.

Motion Restoration. For motion restoration approaches, rotation-based methods achieve perfect skeleton reconstruction (skeleton error: 0.00) but significantly increase jitter error (30.724) and skating performance (0.193). In contrast, our MRO technique also achieves perfect skeleton reconstruction (skeleton error: 0.00) while maintaining low jitter error (2.084) and better skating ratio (0.090). This demonstrates MRO's effectiveness in preserving physical plausibility while maintaining motion quality across different skeletal structures.

Our ablation study confirms each component's critical contribution to the overall system: K-GCA and RVQ enhance cross-skeleton reconstruction and semantic alignment, while MRO ensures physical consistency in the generated motions. These findings validate our architectural design choices for skeleton-aware motion generation and demonstrate SR-VAE's ability to effectively handle arbitrary skeleton motion retargeting with high fidelity.

4.3. Qualitative and Quantitative Analysis

Quantitative Evaluation. We evaluate our methods against state-of-the-art text-to-motion approaches as shown in Table 2. The fundamental distinction of our approach is the ability to directly generate motions for arbitrary skeleton structures—a paradigm shift from existing methods that are restricted to fixed, pre-defined skeletons.

Datasets	Methods	FID↓	R Precision↑		
			Top 1	Top 2	Top 3
Human-ML3D	TM2T [17]	$1.501^{\pm.017}$	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$
	T2M [16]	$1.087^{\pm.021}$	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$
	MDM [41]	$0.544^{\pm.044}$	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$
	MLD [10]	$0.473^{\pm.013}$	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$
	MotionDiffuse [51]	$0.630^{\pm.001}$	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$
	T2M-GPT [50]	$0.141^{\pm.005}$	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$
	ReMoDiffuse [52]	$0.103^{\pm.004}$	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$
	AttT2M [54]	$0.112^{\pm.006}$	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$
	Momask [18]	$0.045^{\pm.002}$	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$
Skele-Motion-3D	**SR-VAE (ours)**	$0.027^{\pm.000}$	$0.508^{\pm.002}$	$0.700^{\pm.002}$	$0.793^{\pm.002}$
	OmniSkel (ours)	$0.084^{\pm.004}$	$0.507^{\pm.002}$	$0.699^{\pm.003}$	$0.797^{\pm.002}$

Table 2. **Comparison with text-to-motion methods on standard skeleton.** Our approach introduces a novel pipeline trained on the SkeleMotion-3D dataset, which fundamentally differs from traditional T2M methods. For fair comparison with existing methods trained on HumanML3D, we configured our models (SR-VAE and OmniSkel) to target the standard HumanML3D skeleton during inference. Despite being designed for arbitrary skeleton generation, our methods achieve competitive performance on this fixed skeleton setting, demonstrating strong text-to-motion capabilities even when constrained to a single skeleton structure.

This fundamental difference presents a challenge for direct comparison, as our methods operate on the newly introduced SkeleMotion-3D dataset with diverse skeleton structures, while previous approaches are confined to the HumanML3D dataset with a single skeleton configuration. Notably, existing methods cannot be trained or evaluated on our dataset due to their inherent architectural limitations in handling skeleton variations.

To provide a meaningful comparison that highlights our text-to-motion capabilities, we configure our models at inference time to target the standard HumanML3D skeleton structure. Despite being designed for the more complex task of arbitrary skeleton generation, both SR-VAE and OmniSkel achieve competitive performance on the HumanML3D test set, with comparable or superior metrics in FID and R-Precision compared to specialized methods.

These results demonstrate that our methods successfully balance adaptability with generation quality, enabling cross-skeleton applications without compromising performance on standard benchmarks.

Qualitative Results. Figure 5 illustrates OmniSkel's capability to generate motions across different skeletal structures. The visualization demonstrates: (1) diverse motions for different textual descriptions, and (2) consistent semantic motion across distinct skeletal anatomies.

Each column represents a different textual prompt, from complex sequences to simpler actions. The model successfully captures the nuanced details of these descriptions, including locomotion styles and sequential actions.

This comparison demonstrates OmniSkel's precise perception and effective integration of different semantics and arbitrary skeletons. More visualization results are provided in the supplementary materials.

4.4. Performance Comparison

Method	Parameters	Inference Time	GPU Memory
MoMask [18]	44.85M	**0.075s**	1434M
Ours	**38.73M**	0.099s	**1402M**

Table 3. Computational cost comparison on RTX 4090.

Our method achieves comparable performance to state-of-the-art approaches while maintaining slight GPU memory advantages as shown in Table 3 and uniquely enabling end-to-end motion generation for arbitrary skeletal specifications. This efficiency demonstrates the practical viability of our skeleton-agnostic framework for real-world applications requiring diverse character animations.

5. Discussion and Conclusion

In conclusion, we present OmniSkel, a text-driven motion generation framework that enables motion synthesis for arbitrary human skeletons through our novel K-GCA and SR-VAE architectures. Our Motion Restoration Optimizer (MRO) addresses the cross-skeleton motion reconstruction challenge, ensuring both motion smoothness and skeletal consistency. Leveraging our SkeleMotion-3D dataset, which pairs textual descriptions with diverse skeletal configurations, our approach bridges the gap between research and practical applications. Experiments demonstrate that OmniSkel achieves competitive performance while maintaining semantic alignment with textual descriptions. This work establishes a foundation for more inclusive motion generation systems that can serve diverse applications without requiring conventional retargeting approaches.

Acknowledgements

This work was supported in part by the the National Natural Science Foundation of China under Grant 62320106007 and Grant 62032022; and in part by the Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

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):62–1, 2020. 2

[2] Adobe. Mixamo. Online Animation Library, 2023. Available at: `https://www.mixamo.com/`. 7

[3] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE, 2018. 2

[4] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 2

[5] André Antakli, Erik Hermann, Ingo Zinnikus, Han Du, and Klaus Fischer. Intelligent distributed human motion simulation in human-robot collaboration environments. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 319–324, 2018. 1

[6] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 2

[7] Gustav Bredell, Kyriakos Flouris, Krishna Chaitanya, Ertunc Erdil, and Ender Konukoglu. Explicitly minimizing the blur error of variational autoencoders. *arXiv preprint arXiv:2304.05939*, 2023. 2

[8] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. 4

[9] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018. 2

[10] 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*, pages 18000–18010, 2023. 1, 2, 8

[11] Kwang-Jin Choi and Hyeong-Seok Ko. Online motion retargetting. *The Journal of Visualization and Computer Animation*, 11(5):223–235, 2000. 2

[12] 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*, 2024. 2

[13] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1396–1406, 2021. 2

[14] Michael Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 33–42, 1998. 2

[15] 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*, pages 2021–2029, 2020. 3

[16] 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*, pages 5152–5161, 2022. 1, 2, 3, 6, 8

[17] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 4, 8

[18] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2, 4, 5, 6, 8

[19] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 2, 4

[20] Moritz Kappel, Vladislav Golyanik, Mohamed Elgharib, Jann-Ole Henningson, Hans-Peter Seidel, Susana Castillo, Christian Theobalt, and Marcus Magnor. High-fidelity neural human motion transfer from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1541–1550, 2021. 1

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

[22] Hema Koppula and Ashutosh Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *International conference on machine learning*, pages 792–800. PMLR, 2013. 1

[23] 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*, pages 39–48, 1999. 2

[24] Sunmin Lee, Taeho Kang, Jungnam Park, Jehee Lee, and Jungdam Won. Same: Skeleton-agnostic motion embedding for character animation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3, 4

[25] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. PM-net: learning of disentangled pose and movement for unsupervised motion retargeting. In *Proc. BMVC*, 2019. 2

[26] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated

videos of human activities from natural language descriptions. *Learning*, 2018(1), 2018. 2

[27] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23222–23231, 2023. 2

[28] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. In *International Conference on Machine Learning*, pages 32939–32977. PMLR, 2024. 2

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 3

[30] Dennis Majoe, Lars Widmer, and Juerg Gutknecht. Enhanced motion interaction for multimedia applications. In *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*, pages 13–19, 2009. 1

[31] 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*, pages 10985–10995, 2021. 2

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 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*, pages 9488–9497, 2023. 2

[34] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. *arXiv preprint arXiv:2403.19435*, 2024. 2

[35] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4

[36] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 2

[37] 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*, pages 8748–8763. PMLR, 2021. 2

[38] 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 2024 Conference Papers*, pages 1–9, 2024. 2

[39] Seyoon Tak and Hyeong-Seok Ko. A physically-based motion retargeting filter. *ACM Trans. Graph.*, 24(1):98–117, 2005. 2

[40] 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*, pages 358–374. Springer, 2022. 2

[41] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 2, 8

[42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[43] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. accepted as poster. 4

[44] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8639–8648, 2018. 2

[45] Congyi Wang. T2m-hifigpt: generating high quality human motion from textual descriptions with residual discrete representations. *arXiv preprint arXiv:2312.10628*, 2023. 2, 4

[46] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12281–12288, 2020. 2

[47] Mohammed Yeasin, Ediz Polat, and Rajeev Sharma. A multiobject tracking framework for interactive multimedia applications. *IEEE transactions on multimedia*, 6(3):398–405, 2004. 1

[48] Haodong Zhang, ZhiKe Chen, Haocheng Xu, Lei Hao, Xiaofei Wu, Songcen Xu, Zhensong Zhang, Yue Wang, and Rong Xiong. Semantics-aware motion retargeting with vision-language models. 2024. 3

[49] Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, and Zhigang Tu. Skinned motion retargeting with residual perception of motion semantics & geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13872, 2023. 2

[50] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 1, 2, 4, 6, 8

[51] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2, 8

[52] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023. 2, 8

[53] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*, 2023. 4

[54] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 509–519, 2023. 2, 8

[55] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. *arXiv preprint arXiv:2403.18512*, 2024. 2