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.

MotionHiFlow: Text-to-Motion via Hierarchical Flow Matching

Heng Li[1] Xiaotong Lin[1] Ling-An Zeng[1] Yulei Kang[1] Shuai Li[2] Jian-Fang Hu[1,3,4†]
[1]Sun Yat-sen University [2]Shandong University
[3] Guangdong Province Key Laboratory of Information Security Technology, China
[4] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{liheng36, linxt29, zenglan3, kangylei}@mail2.sysu.edu.cn, shuaili@sdu.edu.cn, hujf5@mail.sysu.edu.cn

Abstract

Text-to-motion generation aims to generate 3D human motions that are tightly aligned with the input text while remaining physically plausible and rich in fine-grained detail. Although recent approaches can produce complex and natural movements, they usually operate at only one temporal scale, which limits both semantic alignment and temporal coherence. Inspired by the fact that complex motions are conceptualized hierarchically rather than at a single temporal scale in the human cognitive system, we propose MotionHiFlow, a hierarchical flow matching framework to generate motion progressively by constructing flow path from low to high temporal scales. The flows at lower scales capture high-level semantics and coarse motion structures, while flows at higher scales refine temporal details. To link the flows across scales, we introduce a novel cross-scale transition process, ensu ring continuity and preserving noise consistency. Furthermore, by integrating a Text-Motion Diffusion Transformer and a topology-aware Motion VAE, MotionHiFlow explicitly models structural dependencies among joints via joint-aware positional encoding and skeletal topology, enabling precise semantic alignment alongside fine-grained motion details. Extensive experiments on HumanML3D and KIT-ML benchmarks demonstrate state-of-the-art performance, with ablation studies confirming the effectiveness of the hierarchical design and key components. Code is available at https://github.com/ai-lh/MotionHiFlow.

1. Introduction

Text-to-motion generation aims to synthesize realistic 3D human motions conditioned on natural language descriptions, with broad applications in virtual reality, character animation, and robotics. The generated motions are expected to be semantically consistent with the input text and physically plausible with fine-grained motion details. Benefiting from advances in generative modeling and powerful sequence learning architectures, recent methods [15, 30, 38, 41, 56, 62] have made notable progress in generating complex and natural human motions at a single temporal scale.

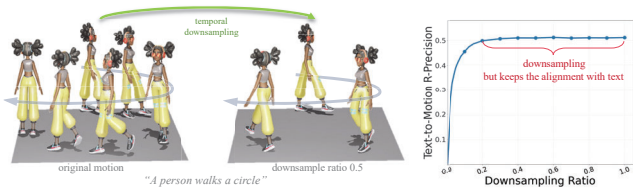

Figure 1. Text-to-Motion retrieval precision under different downsampling ratios. The R-precision remains stable as the downsampling ratio decrease, which means that models trained on coarse motions can achieve robust semantic alignment.

In the human cognitive system [36], complex motions are conceptualized hierarchically rather than at a single temporal scale. Humans typically realize human motions by first constructing a high-level framework of key poses (termed as coarse motion), and subsequently refine it with dynamic transitions and fine-grained limb movements to produce coherent fine motion. However, different from such coarse-to-fine cognitive process, recent methods [15, 41, 62] simultaneously model semantic alignment and motion details at a single temporal scale, which limits their ability to achieve long-term coherence, naturalness, and precise alignment with textual input. To address this gap, we aim to design a hierarchical coarse-to-fine generation strategy that first produces a coarse motion to capture high-level semantic structure at a low temporal scale, and then progressively refines it by adding fine-grained motion details at higher temporal scales.

To verify this intuition, we first evaluate how much semantic information is preserved when motion is viewed at a lower temporal scale (Figure 1). Starting from the original motion, we generate coarse motions by linearly downsampling and evaluate text–motion alignment with the R-precision metric on the HumanML3D [13] test set. Remarkably, R-precision remains stable as the downsampling ratio decreases even at

† Corresponding Author

0.2× (i.e., retaining only 20% of the frames), indicating that coarse motions preserve most of the semantics described in the text. Moreover, by training motion generation models at different scales (refer to Table 2), we observe that models trained solely on coarse motions often achieve robust semantic alignment, sometimes even outperforming those trained on fine-scale motions. These observations suggest that overemphasizing fine-grained details may hinder semantic learning, while training on coarse motions promotes stronger alignment with the core textual semantics.

Building on the above observations, we propose a hierarchical framework *MotionHiFlow* to generate motion progressively from low to high temporal scales across multiple stages, aiming to achieve both strong semantic alignment and rich motion details. Specifically, we devise a stage-wise flow to link the start (noisier latent) and end (cleaner latent) at each scale. Instead of directly upsampling lower-scale noisy data as done in other works [3, 22], here we formulate a novel cross-scale transition process to link flows across scales, which contains: 1) denoising: constructing the clean data at the lower scale by extrapolation; 2) upsampling: generating clean data at higher scale with upsampling; 3) renoising: constructing the noise data at the higher scale via interpolation. We define the transition in this way such that noise consistency can be preserved across stages. By integrating the cross-scale transition with stage-wise flows, a generative process is established that maps noise to data. To operationalize MotionHiFlow, we further introduce *Text-Motion Diffusion Transformer (TMDiT)*, a novel motion generation model that harnesses hierarchical flow matching for efficient and smooth motion generation. Built upon recent success of diffusion models in image generation [11, 27, 39], TMDiT explicitly incorporates the inherent structural dependencies among human joints through joint-aware positional encoding (Joint RoPE). By combining TMDiT with a topology-aware Motion VAE which encodes motion sequence into latent, our model effectively generates motions with both semantic alignment and fine-grained details.

We conduct quantitative and qualitative comparisons together with extensive ablation studies on the HumanML3D [13] and KIT-ML [44] datasets to verify the effectiveness of our hierarchical design and each model component. In summary, our main contributions are as follows:

- We propose *MotionHiFlow*, a hierarchical flow matching framework for text-to-motion generation that progressively and consistently generates motion from low to high temporal scales, achieving strong semantic alignment and rich fine-grained motion details.
- We develop the *Text-Motion Diffusion Transformer* (TMDiT), which explicitly incorporates the inherent structural dependencies among human joints through joint-aware positional encoding and skeletal topology.
- We achieve state-of-the-art performance on the HumanML3D and KIT-ML datasets, with comprehensive ablations validating the effectiveness of our approach.

2. Related work

2.1. Text-Conditioned Motion Generation

Text-conditioned human motion generation is a challenging task. Early attempts [1, 13, 40] primarily focused on learning direct mappings or shared embeddings between text descriptions and motions. For instance, Language2Pose [1] proposed learning a common latent representation subspace for both modalities. Despite significant progress, these methods still struggle to generate diverse and high-fidelity motions. More recently, diffusion-like models [4, 7, 8, 28, 42, 46, 47, 53, 57, 62, 64, 69, 72] have become the dominant approach for text-conditioned human motion generation, significantly advancing the field. For example, MoGenTS [62] builds on this by enhancing the spatial VAE component for improved motion representation. Furthermore, other methods include autoregressive approaches [2, 14, 20, 66, 73], generative masked modeling techniques [15, 41, 70], and additional types [9, 19, 29, 33–35, 50–52, 58, 59, 63, 71, 74]. Moreover, some works [9, 21, 23, 24, 26, 43, 55, 60, 65] explore using supplementary conditions to control the generated motion. However, these models operate at a single temporal scale, which limits their ability to simultaneously capture global trajectory structures (requiring coarse temporal views) and fine-grained motion details (necessitating fine temporal views). In contrast, we develop a hierarchical flow matching framework that enables multi-scale generation, starting with high-level semantic alignment to the text and progressively refining motion details in a top-down manner.

2.2. Flow Generative Models

Flow Matching, a powerful generative modeling paradigm [31, 32], has recently achieved compelling results in domains including image and video generation [11, 22, 27]. By leveraging theoretical advantages such as straight flow formulations and direct conditional path probability learning, flow generative models offer benefits in training stability and efficient generation compared to diffusion-based models. For example, Patrick et al. [11] propose to learn a continuous path between a high-dimensional simple noise distribution and the target high-resolution image manifold. However, flow generative models in 3D human motion generation remain largely underexplored. Although a few methods have made initial attempts [6, 18], they naively apply flow matching for generation in the motion space without adaptation, resulting in suboptimal utilization and generation performance. To address this, we enhance flow matching by integrating Text-Motion Diffusion Transformer with a joint-aware positional encoding, fully unlocking the significant potential of flow matching models for human motion generation.

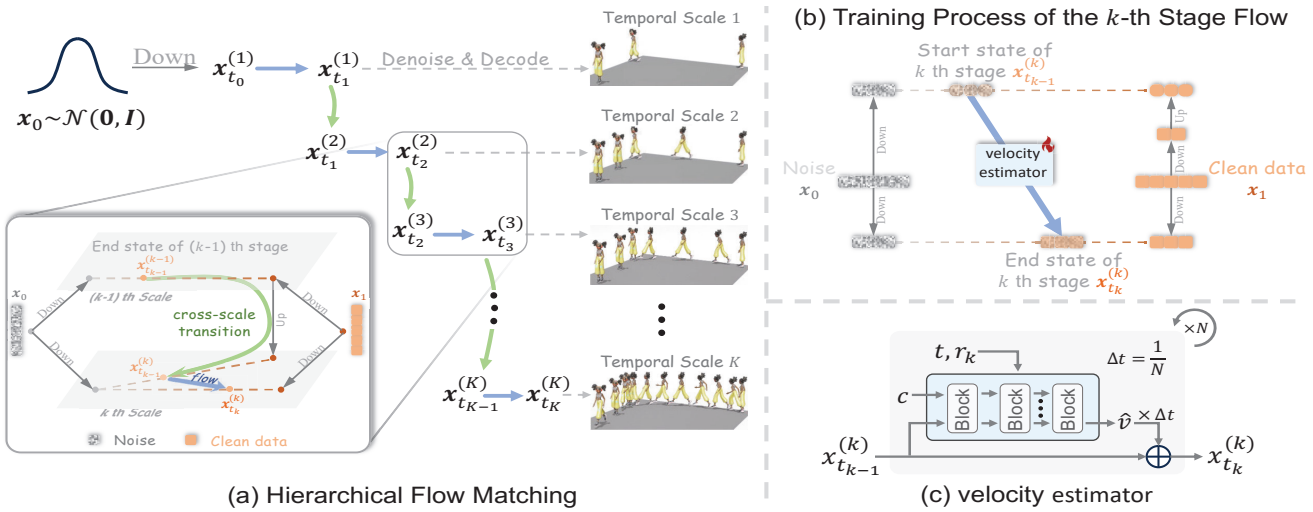

Figure 2. Overview of our MotionHiFlow, which progressively generates motion from low to high temporal scales across multiple stages. The early stages mainly capture high-level semantics and coarse motion structures, while later stages model fine-grained temporal details via *cross-scale transition* and *flow* operators. The points along the gradient-colored dashed line in the inset (bottom left) of (a) and in (b) denotes a linear interpolation between its endpoints. Down/Up denotes downsampling/upsampling, respectively.

3. Method

3.1. Problem Formulation

Given a text query describing human motion or action, our goal is to generate a corresponding 3D human pose sequence $M = \{m_i\}$ of length L. According to existing works [13, 15], each joint in the 3D human pose is represented by the root angular velocity along the Y-axis, root linear velocities on the XZ plane, root height, and local joint positions, rotations and velocities relative to the root space. Please refer to T2M [13] for more detailed information on human pose representation. Each human pose is represented as a $J \times D_j$-shaped tensor, where J indicates the joint number, and D_j is the dimension of joint representation.

3.2. Overview

To generate semantically aligned, temporally coherent, and detailed motion, we propose a novel hierarchical flow matching framework, MotionHiFlow. This framework operates in the latent space encoded by a Motion VAE that provides topology-aware encoding of motion sequences. As shown in Figure 2, our framework progressively generates the motion from low to high temporal scales across multiple stages. The early stages mainly capture high-level semantics and coarse motion structures, while later stages add fine-grained temporal details. To facilitate the continuity flow matching across stages, we formulate a novel cross-scale transition with denoising-upsampling-renoising process, preserving noise consistency over the entire motion generation process. By integrating the cross-scale transition with stage-wise flows, a generative process is established that maps noise to data. Other key components include a Motion VAE with a two-stream Graph Convolutional Networks [48] for motion tokenization, a Text-Motion Diffusion Transformer (TMDiT) modeling hierarchical flow conditioned on text, and a Joint RoPE mechanism for improved positional encoding. Combining with these components, the clean latents outputted by our MotionHiFlow can be decoded to realistic, expressive motions that closely match textual descriptions.

3.3. Preliminaries: Flow Matching

Flow matching models [11, 31, 32] aim to learn a velocity field u_t that transforms noise $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ into data samples $\boldsymbol{x}_1 \sim q_{\text{data}}$. To achieve this, a neural network $v_\theta(\boldsymbol{x}_t, t)$ parameterized by θ, is trained to approximate the target field $u_t(\boldsymbol{x}_t|\boldsymbol{x}_1)$ by minimizing the following loss:

$$\mathcal{L}_{FM}(\theta) = \mathop{\mathbb{E}}_{\substack{t\sim\mathcal{U}(0,1)\\ p_t(\boldsymbol{x}_t|\boldsymbol{x}_1),q(\boldsymbol{x}_1)}} \|v_\theta(\boldsymbol{x}_t, t) - u_t(\boldsymbol{x}_t|\boldsymbol{x}_1)\|^2. \quad (1)$$

Here, $p_t(\boldsymbol{x}_t|\boldsymbol{x}_1)$ defines the conditional probability path linking intermediate point $\boldsymbol{x}_t$ to $\boldsymbol{x}_1$. A common choice [31] for $\boldsymbol{x}_t$ at this path is linear interpolation between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1. \quad (2)$$

Then, the corresponding target velocity field simplifies to $u_t(\boldsymbol{x}_t|\boldsymbol{x}_1) = \boldsymbol{x}_1 - \boldsymbol{x}_0$. The new data can be constructed by solving the following Ordinary Differential Equation (ODE) with existing solvers (e.g., Euler, RK45 [10]):

$$\mathrm{d}\boldsymbol{x}_t = v_\theta(\boldsymbol{x}_t, t)\mathrm{d}t. \quad (3)$$

3.4. Hierarchical Flow Matching Framework

We propose a hierarchical flow matching framework for the text-to-motion generation task, which generates motion progressively from low to high temporal scales across multiple

stages, aiming to achieve both strong semantic alignment, high temporally coherent and rich detailed motion. As illustrated in Figure 2 (a), our framework contains K stages of generation, each operates at a certain temporal scales. Early stages focus on the semantics of coarse motion and text alignment, and the subsequent stages intend to progressively enrich the details of the motion provided by preceding stage.

Specifically, in the k-th stage, our method processes motion at temporal scale $r_k \in (0, 1]$ and refines the motion representation within time interval $[t_{k-1}, t_k]$, where $\{t_k\}_{k=0}^{K}$ are time points partitioning the interval $[0, 1]$. We utilize flow matching to learn a flow transformation S_k that maps the start state $\boldsymbol{x}_{t_{k-1}}^{(k)}$ to end state $\boldsymbol{x}_{t_k}^{(k)}$, as defined below:

$$\text{Start:} \quad \begin{aligned} \boldsymbol{x}_{t_{k-1}}^{(k)} &= (1 - t_{k-1}) f(\boldsymbol{x}_0, r_k) \\ &\quad + t_{k-1} f(f(\boldsymbol{x}_1, r_{k-1}), r_k / r_{k-1}) \end{aligned}, \tag{4}$$

$$\text{End:} \quad \boldsymbol{x}_{t_k}^{(k)} = (1 - t_k) f(\boldsymbol{x}_0, r_k) + t_k f(\boldsymbol{x}_1, r_k). \tag{5}$$

Here, $f(\boldsymbol{x}, r)$ means performing a temporal resampling on $\boldsymbol{x}$ with a factor r, it is downsampling when $r \in (0, 1)$ and upsampling when $r > 1$. The end state $\boldsymbol{x}_{t_k}^{(k)}$ at scale r_k is defined as a linear interpolation between noise $\boldsymbol{x}_0$ and clean data $\boldsymbol{x}_1$. The start state $\boldsymbol{x}_{t_{k-1}}^{(k)}$ is carefully designed to incorporate the information $f(\boldsymbol{x}_1, r_{k-1})$ from the previous stage and initial noise $\boldsymbol{x}_0$, maintaining noise consistency across stages. By integrating the flow S_k across stages, we define the generative path of MotionHiFlow from noise to data, which can be trained by minimizing the following loss:

$$\mathcal{L}_{HFM}(\theta) = \mathop{\mathbb{E}}_{k,t} \left\| v_\theta(\boldsymbol{x}_t^{(k)}, t) - (\boldsymbol{x}_{t_k}^{(k)} - \boldsymbol{x}_{t_{k-1}}^{(k)}) \right\|^2. \tag{6}$$

Here, $\boldsymbol{x}_t^{(k)}$ represents training points sampled in k th stage, which is defined as:

$$\boldsymbol{x}_t^{(k)} = (1 - \tau)\boldsymbol{x}_{t_{k-1}}^{(k)} + \tau \boldsymbol{x}_{t_k}^{(k)}, \tag{7}$$

where $\tau = (t - t_{k-1})/(t_k - t_{k-1})$ is the normalized time within the stage k.

During inference, instead of directly upsampling lower-scale noisy data as done in other works [3, 22], which leads to noise inconsistency and thus degrades generation performance. Here we formulate a novel cross-scale transition process to bridge the start state of stage $k + 1$ with the end state of stage k across scales. This process involves three steps: 1) denoise: constructing the clean data at the lower scale by extrapolation; 2) upsample: generating clean data at the higher scale with upsampling; 3) renoise: constructing the noise data at the higher scale via interpolation. Formally, the transition is governed by the following equations:

$$\text{denoise:} \quad \hat{\boldsymbol{x}}_1^{(k)} = \left[\hat{\boldsymbol{x}}_{t_k}^{(k)} - (1 - t_k)\boldsymbol{x}_0^{(k)} \right] / t_k, \tag{8}$$

$$\text{upsample:} \quad \hat{\boldsymbol{x}}_1^{\prime(k+1)} = f(\hat{\boldsymbol{x}}_1^{(k)}, r_{k+1}/r_k), \tag{9}$$

$$\text{renoise:} \quad \hat{\boldsymbol{x}}_{t_k}^{(k+1)} = (1 - t_k)\boldsymbol{x}_0^{(k+1)} + t_k \hat{\boldsymbol{x}}_1^{\prime(k+1)}. \tag{10}$$

Algorithm 1 Hierarchical Flow Matching Inference

Require: Trained model parameters θ (for v_θ), scale schedule $r_1, \ldots, r_K$, time partition intervals $[t_0, t_1, \ldots, t_{K-1}, t_K]$ with $t_0 = 0, t_K = 1$, temporal linear resampling function $f(\cdot, \cdot)$, prior distribution for noise (e.g., $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$), conditioning c.

Ensure: Generated clean sample $\hat{\boldsymbol{x}}_1$ at full scale.

```
Sample the initial noise x_0 ~ N(0, I)
Initialize the start state: x̂_0^(k) ← f(x_0, r_1)
for k = 1 to K do
    Flow from x̂_{t_{k-1}}^(k) to x̂_{t_k}^(k)                ▷ Equation (3)
    Cross-scale transition, get the start state of the next scale x̂_{t_k}^(k+1)   ▷ Equation (8), (9) and (10)
end for
x̂_1 ← f(x̂_{t_K}^(K), 1/r_K)                                ▷ r_K may be smaller than 1
return x̂_1
```

Here, $\boldsymbol{x}_0^{(k)}$ denotes $f(\boldsymbol{x}_0, r_k)$, and similarly for $\boldsymbol{x}_0^{(k+1)}$. This three-step process preserves noise consistency and ensures smooth transitions across scales, thus facilitating robust inference across all stages. The complete inference process is presented in Algorithm 1.

By integrating the cross-scale transition with stage-wise flows, this generative process constitutes a deterministic ODE trajectory from the initial noise $\boldsymbol{x}_0$ towards the data distribution, avoiding the need to introduce additional noise between stages as done in related methods [3, 22].

3.5. Model Architecture

For efficient and stable training, we introduce a specific model architecture comprising several key elements. First, we employ a Motion VAE to project input motion into a topology-aware latent space and then reconstruct the motions from latent representation. Then, we develop the Text-Motion Diffusion Transformer (TMDiT), which explicitly incorporates the inherent structural dependencies among human joints through joint-aware positional encoding (Joint RoPE). These components combined with MotionHiFlow form a strong text-to-motion generation model.

Motion VAE. Our Motion VAE is utilized to encode the input motion $M \in \mathcal{R}^{L \times J \times D_j}$ into a compact latent representation $\boldsymbol{x} \in \mathcal{R}^{l \times j \times d}$ and reconstruct the original motion sequence from this compact latent representation. Here, L, J, and D_j denote the number of frames, the number of input joints per frame, and the dimension of joint features (e.g., rotation, velocity), respectively. Correspondingly, l is the latent temporal length and j is the number of latent joints. Unlike MoGenTS [62], which primarily employs 2D convolutions, our Motion VAE utilizes Graph Convolutional Networks (GCNs) [17, 48, 61] to explicitly capture human body topology. The encoder performs temporal downsampling by a

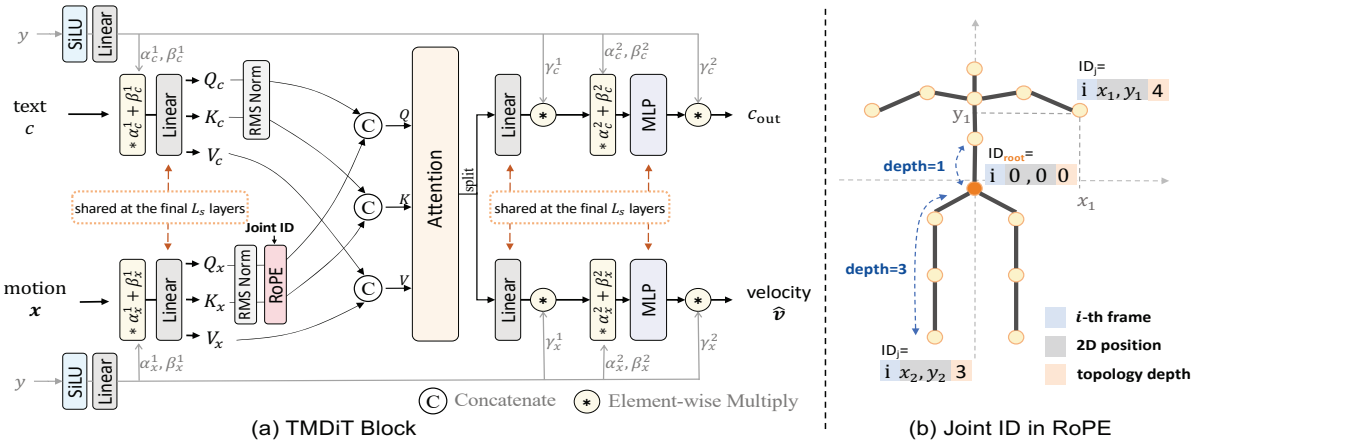

Figure 3. Illustration of two main components in our TMDiT. (a) The TMDiT block employs two separate streams that independently process motion and text features, while self-attention and shared parameters enables information exchange between streams. (b) The Joint RoPE integrates rotations derived from temporal displacement, relative spatial coordinates, and the human body topologies. Here, x_j and y_j denote the j-th joint's coordinates in the xy-plane.

factor 4 (resulting in $l = \lfloor L/4 \rfloor$)), and spatial graph downsampling (from J to j latent joints) using methods such as averaging or learnable pooling [61]. This topology-aware encoding yields modest improvements in reconstruction quality, it notably enhances generation quality when combined with our TMDiT and Joint RoPE.

Text-to-Motion Diffusion Transformer (TMDiT). We present the Text-to-Motion Diffusion Transformer (TMDiT), a new architecture for generating human motion conditioned on textual descriptions. Our approach is motivated by recent progress in diffusion transformers, notably MMDiT [11] and Flux [27]. Crucially, TMDiT departs from conventional methods (e.g., [15, 42, 62]) that typically adapt vanilla Transformers [54] by processing motion sequences alongside a single, sentence-level text embedding c_{vec}. We would like to point out that this prior strategy limits the nuanced interplay required between text and motion. As illustrated in Figure 2 (b), TMDiT receives noised motion latent $\boldsymbol{x}_t^{(k)}$and conditioning word-level text embedding $\boldsymbol{c}$ encoded by CLIP [45]. Information about the current timestep t, sentence-level text embedding c_{vec} and staged scale r_k is fused into an embedding $\boldsymbol{y}$, which then modulates the TMDiT blocks [11]. The output corresponding to the motion component serves as the estimated velocity field for the underlying ODE solver.

Within a TMDiT block, as shown in Figure 3 (a), we employ separate processing streams for the motion features $\boldsymbol{x}$ and text features $\boldsymbol{c}$. Both $\boldsymbol{x}$ and $\boldsymbol{c}$ are separately fed into distinct linear transformations before and after the attention, as well as within the feedforward MLP. The conditioning embedding $\boldsymbol{y}$ modulates these operations through scaling and shifting (pre-attention and pre-MLP) and gating (post-attention and post-MLP), akin to established practices in diffusion models [11, 39]. For clarity of presentation in Figure 3 (a), LayerNorm operations are not explicitly shown, which precede the scaling/shifting. To effectively capture shared representations while preserving modality-specific nuances, TMDiT employs a parameter sharing scheme [27]. The early layers of the network utilize separate parameters for motion and text pathways, allowing for independent feature extraction. Conversely, the final L_s layers share parameters. It first refines modality-specific characteristics and then encourages the learning of common representations essential for coherent text-to-motion generation.

Joint RoPE. We introduce Joint RoPE, an adaptation of Rotary Position Embedding [49] optimized for skeletal motion generation. Joint RoPE encodes relative positions by integrating rotations derived from temporal displacement, relative spatial coordinates, and the kinematic tree structure. A key innovation is its enforcement of skeletal symmetry, ensuring identical relative rotations for symmetrically equivalent joint pairs (e.g., left-hand to right-hand vs. left-foot to right-foot) at the same temporal offset, embedding structural bias akin to RoPE's temporal encoding. Within each attention head, feature dimensions are divided into four segments with proportions $[1/2, 1/8, 1/8, 1/4]$, each applying a 1D RoPE. The first segment $(1/2)$ encodes the joint's temporal position in the motion sequence. The next two segments $(1/8$ each, totaling $1/4)$ encode the joint's 2D spatial coordinates relative to the pelvis in a reference T-pose. The final segment $(1/4)$ encodes the joint's depth in the kinematic tree, with the pelvis as the root. Temporal indices are scaled by a factor r_k before applying RoPE, enabling the TMDiT to generate spatially and structurally coherent motions. By unifying all positional encodings under RoPE framework, Joint RoPE can yield performance improvements, offer potential scalability to varying joint counts, and enhance adaptability across diverse skeletal structures.

3.6. Training and Inference

Two-Stage Training. Our training pipeline has two stages. In the first stage, we train the Motion VAE by minimizing a composite objective that combines the standard VAE losses (reconstruction and KL loss) and an auxiliary term that improves the temporal robustness of the latent representation. For a random subset of each batch, we downsample the latent vector $\boldsymbol{x}$ by a factor $r \in [0.3, 1]$ with a linear resampling function f, then compute the mean-squared error (MSE) between the Motion VAE decoder output (generated from the downsampled latent) and the corresponding downsampled motion sequence M:

$$\mathcal{L}_{\text{aug}} = \|\text{Dec}(f(x, r)) - f(M, r)\|^2. \tag{11}$$

In the second stage, we freeze the Motion VAE and train the TMDiT model v_θ using the hierarchical flow-matching loss (Eq. 6). To enable classifier-free guidance (CFG) [16], we apply condition dropout during training, randomly replacing the text condition c with a null token $\varnothing$ at 10 % probability. Training is conducted following the standard flow matching paradigm for each scale r_k, with the timestep t uniformly sampled from the interval $[t_{k-1},\, t_k]$.

Inference. Given the well-trained model parameters v_θ and random noise $\boldsymbol{x}_0$, we generate the motion latent representations $\hat{\boldsymbol{x}}_1$ following the procedure presented in Algorithm 1. Specifically, we enhance the velocity estimation with CFG [16]. The resulting latent $\hat{\boldsymbol{x}}_1$ is finally inputted to the Motion VAE decoder, producing the motion sequence.

4. Experiments

In this section, we conduct extensive experiments on two widely used benchmark datasets. The results suggest that our MotionHiFlow consistently outperforms the current state-of-the-art methods quantitatively and qualitatively. Furthermore, ablation studies demonstrate the effectiveness of the key designs in our proposed framework.

4.1. Experimental Setup

Datasets. We evaluate our MotionHiFlow framework on the widely used benchmark datasets: HumanML3D dataset [13] and KIT-ML dataset [44]. The HumanML3D dataset [13] comprises $14,616$ human motions derived from the AMASS [37] and HumanAct12 [12] collections. Each motion is annotated with three distinct textual descriptions, resulting in $44,970$ motion-text pairs in total. The KIT-ML dataset [44] contains $3,911$ motions paired with $6,278$ textual descriptions. Both datasets are split into training 80%, validation 5%, and test 15% sets.

Evaluation Metrics. We employ the same evaluation configuration as previous works [13, 15, 66, 69]. To evaluate the semantic alignment between generated motions and input texts, we adopt *R-Precision* and *Multimodal Distance* as metrics, which is computed as the top-k recall precision and the multimodal distance between generated motions and input texts, respectively. We also employ the *Fréchet Inception Distance (FID)* to measure the feature distributional similarity on the latent feature space between ground-truth (GT) and generated motions. Additionally, the *Diversity* measures the variance across generated motion sequences, computed as the average Euclidean distance between 300 randomly sampled motion pairs. However, this metric typically shows similar values across methods and is thus less emphasized. Note that all metrics are calculated using pretrained a text encoder and motion encoder from T2M [13].

Implementation Details. The Graph Convolutional Network (GCN) in our Motion VAE is adapted from 2s-AGCN [48]. The encoder comprises two blocks, each integrating GCN and Temporal Convolutional Network (TCN) layers. This reduces the temporal dimension by a factor of 4 (resulting in a sequence length of $L/4$) and pools the skeleton graph into $j = 6$ latent joints, representing the torso, pelvis, and four limbs (arms and legs). The weight of the augmentation loss (Eq. 11) is empirically set to 0.5. For TMDiT, we employ nine blocks, with the first three using distinct parameters for the dual branches and the latter six sharing parameters. The latent dimension is set to 384, with 6 attention heads and a feed-forward dimension of 1536. The hierarchical flow matching comprises three flow layers, operating at scales $r_k \in \{1/3, 2/3, 1\}$.

The MotionVAE is trained for 300,000 steps using AdamW [25] with a batch size of 256 and an initial learning rate of 2×10^{-4}. The TMDiT is subsequently trained for 200,000 steps using AdamW, with a batch size of 64 and an initial learning rate of 2×10^{-4}. For both models, we apply a MultiStepLR scheduler, reducing the learning rate by a factor of 0.2 at 50% and 75% of the total training steps.

4.2. Comparison with State-of-the-art Methods

We compare our MotionHiFlow with various state-of-the-art methods, including VAE-based approaches [40], autoregressive models [29, 66], diffusion-based models [17, 64, 67–69] and discrete diffusion-like [15, 62] models.

Quantitative Results. Table 1 presents the quantitative comparisons between our MotionHiFlow and existing methods on the HumanML3D [13] and KIT-ML [44] datasets. Each experiment is repeated 20 times, with results reported along with a 95% statistical confidence interval to ensure reliability. Table 1 shows that MotionHiFlow outperforms every baseline on both HumanML3D and KIT-ML. It achieves the highest R-Precision (0.563 / 0.482) and the lowest FID (0.032 / 0.135), evidencing tighter text–motion alignment and more realistic motion. Simultaneously, it records the smallest MultiModal Distance, indicating outputs that are both semantically coherent and varied. These results validate

Table 1. Quantitative comparisons with the current state-of-the-art methods on the HumanML3D (upper half) and KIT-ML (lower half) datasets. Symbol "±" denotes a 95% confidence interval. Text in **bold** and underline denote the best and second-best results, respectively.

Methods	Venue	R-Precision ↑			FID↓	MultiModal Dist ↓	Diversity →
		Top1	Top2	Top3			
On the HumanML3D dataset [13].							
TEMOS [40]	*ECCV*'22	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$
T2M-GPT [66]	*CVPR*'23	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$9.761^{\pm.081}$
ReMoDiffuse [68]	*ICCV*'23	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$9.018^{\pm.075}$
MoMask [15]	*CVPR*'24	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	-
BAMM [41]	*ECCV*'24	$0.525^{\pm.002}$	$0.720^{\pm.003}$	$0.814^{\pm.003}$	$0.055^{\pm.002}$	$2.919^{\pm.008}$	$9.717^{\pm.089}$
MoGenTS [62]	*NeurIPS*'24	$0.529^{\pm.003}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$\underline{0.033}^{\pm.001}$	$2.867^{\pm.006}$	$9.570^{\pm.077}$
Light-T2M [64]	*AAAI*'25	$0.511^{\pm.003}$	$0.699^{\pm.002}$	$0.795^{\pm.002}$	$0.040^{\pm.002}$	$3.002^{\pm.008}$	-
IRG-MotionLLM [29]	*arXiv*'25	$0.535^{\pm.002}$	$0.725^{\pm.002}$	$0.820^{\pm.002}$	$0.242^{\pm.006}$	$2.785^{\pm.006}$	$9.900^{\pm.094}$
EnergyMoGen [67]	*CVPR*'25	$0.526^{\pm.003}$	$0.718^{\pm.003}$	$0.815^{\pm.002}$	$0.176^{\pm.006}$	$2.931^{\pm.007}$	$9.500^{\pm.091}$
SALAD [17]	*CVPR*'25	$\mathbf{0.581}^{\pm.003}$	$\mathbf{0.769}^{\pm.003}$	$\mathbf{0.857}^{\pm.002}$	$0.076^{\pm.002}$	$\mathbf{2.649}^{\pm.009}$	$9.696^{\pm.096}$
MoMask++ [5]	*NeurIPS*'25	$0.528^{\pm.003}$	$0.718^{\pm.003}$	$0.811^{\pm.002}$	$0.072^{\pm.003}$	$2.912^{\pm.008}$	-
MotionHiFlow (ours)	-	$\underline{0.563}^{\pm.003}$	$\underline{0.754}^{\pm.003}$	$\underline{0.843}^{\pm.003}$	$\mathbf{0.032}^{\pm.002}$	$\underline{2.691}^{\pm.009}$	$9.504^{\pm.071}$
On the KIT-ML dataset [44].							
TEMOS [40]	*ECCV*'22	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.717^{\pm.051}$	$3.417^{\pm.019}$	$10.84^{\pm.100}$
T2M-GPT [66]	*CVPR*'23	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.86^{\pm.094}$
ReMoDiffuse [68]	*ICCV*'23	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$0.155^{\pm.006}$	$2.814^{\pm.012}$	$10.80^{\pm.105}$
MoMask [15]	*CVPR*'24	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	$2.779^{\pm.022}$	-
BAMM [41]	*ECCV*'24	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$0.788^{\pm.005}$	$0.183^{\pm.013}$	$2.723^{\pm.026}$	$11.008^{\pm.094}$
MoGenTS [62]	*NeurIPS*'24	$0.445^{\pm.006}$	$0.671^{\pm.006}$	$0.797^{\pm.005}$	$\underline{0.143}^{\pm.004}$	$2.711^{\pm.024}$	$10.918^{\pm.090}$
Light-T2M [64]	*AAAI*'25	$0.444^{\pm.006}$	$0.670^{\pm.007}$	$0.794^{\pm.005}$	$0.161^{\pm.009}$	$2.746^{\pm.016}$	-
IRG-MotionLLM [29]	*arXiv*'25	$0.445^{\pm.005}$	$0.681^{\pm.003}$	$0.781^{\pm.004}$	$0.432^{\pm.013}$	$2.740^{\pm.017}$	$11.115^{\pm.086}$
EnergyMoGen [67]	*CVPR*'25	$0.436^{\pm.006}$	$0.651^{\pm.006}$	$0.772^{\pm.006}$	$0.495^{\pm.020}$	$2.861^{\pm.020}$	$11.06^{\pm.101}$
SALAD [17]	*CVPR*'25	$\underline{0.477}^{\pm.006}$	$\mathbf{0.711}^{\pm.005}$	$\mathbf{0.828}^{\pm.005}$	$0.296^{\pm.012}$	$\underline{2.585}^{\pm.016}$	$11.097^{\pm.095}$
MotionHiFlow (ours)	-	$\mathbf{0.482}^{\pm.005}$	$\underline{0.704}^{\pm.005}$	$\underline{0.825}^{\pm.005}$	$\mathbf{0.135}^{\pm.007}$	$\mathbf{2.552}^{\pm.014}$	$10.894^{\pm.117}$

that our hierarchical flow matching, joint-aware encoding, and diffusion transformer jointly set a new state of the art.

Qualitative Results. Figure 4 presents qualitative comparisons of our MotionHiFlow against prior methods, including Momask [15], BAMM [41], and MoGenTS [62]. As shown, existing approaches often misinterpret directional cues and specific limb movements. In contrast, MotionHiFlow produces motions with enhanced dynamic realism and strong alignment to textual descriptions, further validating the effectiveness of our hierarchical flow matching framework.

4.3. Ablation Study

In this subsection, we conduct ablation studies on the HumanML3D [13] dataset to comprehensively analyze the influence of different components in our MotionHiFlow, including the hierarchical flow matching and the architecture.

Effectiveness of Hierarchical Flow Matching. In Table 2, we investigate the effect of hierarchical designs in our MotionHiFlow framework. Specifically, we test the system performance under varying numbers of hierarchical scales and different temporal scales. As can be seen: 1) even trained completely in a single coarse scale, our system can obtain MM-Dist ranging from 2.717 ($\{r_k\} = [0.4]$) to 2.729 ($\{r_k\} = [0.6]$), which means that a good semantic alignment is achieved; 2) performing hierarchical flow matching can improve both the FID and MM-Dist metrics.

Table 2. Evaluation of system performance under varying scale settings. R@1 denotes top-1 retrieval precision.

scales $\{r_k\}$	FID ↓	R@1 ↑	MM-Dist ↓
$[0.4]$	$0.106^{\pm.004}$	$0.561^{\pm.003}$	$2.717^{\pm.006}$
$[0.6]$	$0.061^{\pm.003}$	$0.556^{\pm.003}$	$2.729^{\pm.006}$
$[0.8]$	$0.058^{\pm.002}$	$0.559^{\pm.003}$	$2.722^{\pm.005}$
$[1]$	$0.051^{\pm.002}$	$0.556^{\pm.003}$	$2.723^{\pm.006}$
$[1/2, 1]$	$0.038^{\pm.001}$	$\mathbf{0.565}^{\pm.002}$	$2.702^{\pm.006}$
$[1/3, 2/3, 1]$	$\mathbf{0.032}^{\pm.002}$	$0.563^{\pm.003}$	$\mathbf{2.691}^{\pm.009}$
$[1/4, 2/4, 3/4, 1]$	$0.035^{\pm.002}$	$0.560^{\pm.003}$	$2.693^{\pm.007}$

Analysis on the Architecture Designs. To evaluate the efficacy of the architectural components in our MotionHiFlow, we begin with the baseline model that employing a standard Transformer Encoder [54] augmented with AdaLN [39]. We

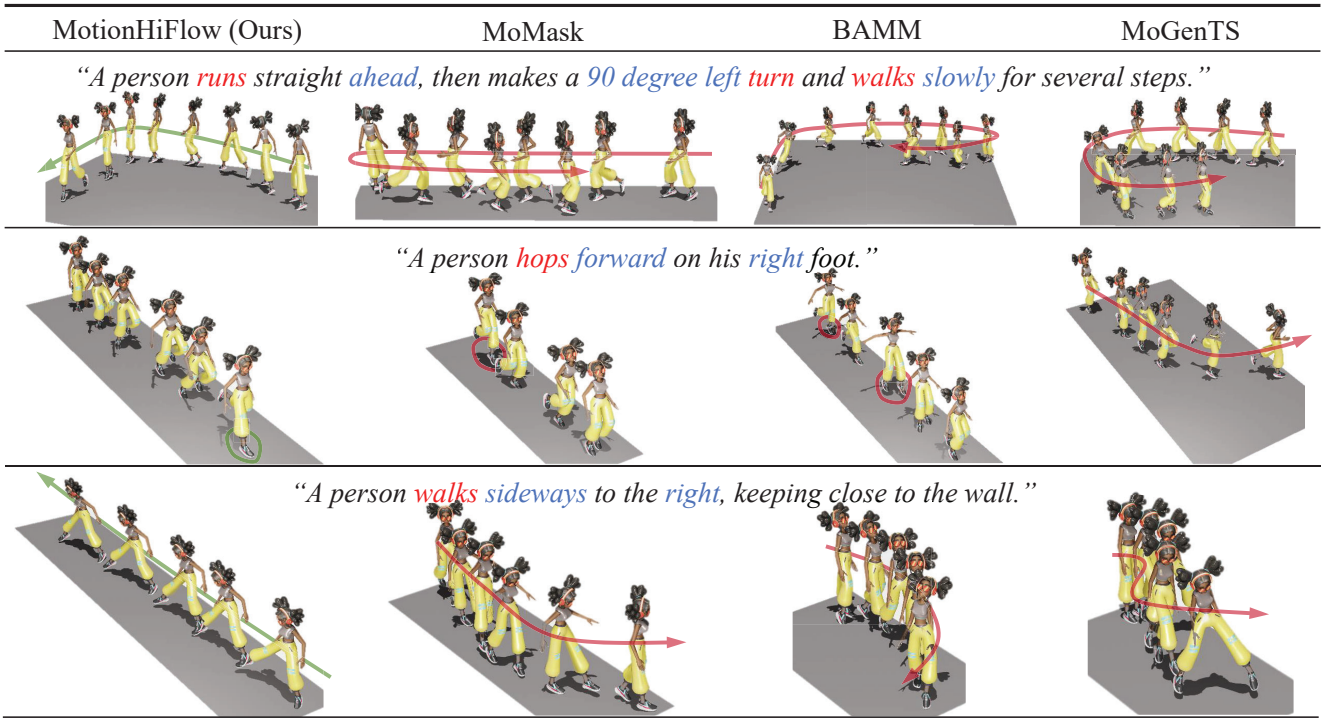

Figure 4. Visual comparisons between different methods given three distinct text descriptions. Only key frames are displayed, with arrows indicating the character's movement direction. Green lines denote correct directions, while red lines indicate incorrect directions that do not match the text content. Refer to the demo video for complete motion clips and more visualization results.

Table 3. Evaluation of key components on the system performance.

	FID ↓	R@1↑	MM-Dist ↓
Baseline	$0.074^{\pm.003}$	$0.511^{\pm.003}$	$3.043^{\cdot 008}$
+ TMDiT	$0.045^{\pm.003}$	$0.557^{\pm.003}$	$2.738^{\pm.007}$
+ topology-aware VAE	$\mathbf{0.032}^{\pm.002}$	$\mathbf{0.563}^{\pm.003}$	$\mathbf{2.691}^{\pm.009}$

incrementally integrate the TMDiT module and the topology-aware Motion VAE. As reported in Table 3, the incorporation of TMDiT yields a substantial improvement in the MM-Dist metric. We hypothesize that this enhancement stems from the word-level text encoding and the non-shared parameter strategy, validated through extensive experiments detailed in the supplementary material. Furthermore, the enhanced Motion VAE introduces spatial information, facilitating finer-grained motion generation.

User Study. We further conduct user study to compare the results generated by MotionHiFlow and previous methods including MoMask [15], MoGenTS [62], and ground truth. We generate 100 motions for each pair of competitors using the text pool provided in HumanML3D test set and present the visualization results side-by-side. A total of 20 users are asked to vote which result is better from the aspect of realism and text alignment, respectively. The detailed results are presented in Figure 5. As shown, MotionHiFlow is preferred

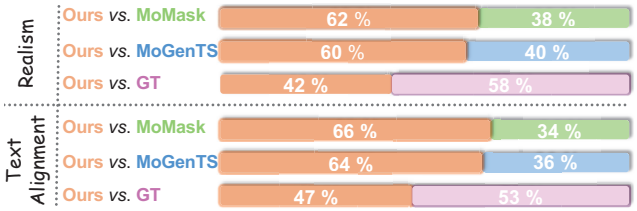

Figure 5. Results of a user study comparing the realism and text alignment of various methods, including our approach, with baseline competitors (MoMask, MoGenTS), and Ground Truth.

by users over MoMask and MoGenTS in both realism and text alignment, even with a 47% chance of surpassing the ground truth in text alignment.

5. Conclusion

In this work, we propose MotionHiFlow, a hierarchical flow matching framework for text-to-motion generation that progressively generates the motion from low to high temporal scales. By integrating a topology-aware Motion VAE and Text-Motion Diffusion Transformer, our method generates motions with more superior semantic alignment, temporal coherence, and dynamic realism as compared with existing approaches, which is experimentally demonstrated on the HumanML3D [13] and KIT-ML [44] datasets.

Acknowledgements

This work was supported partially by the NSFC (62476296), Guangdong Natural Science Funds Project (2023B1515040025, 2022B1111010002, 2024A1111120017), Guangdong NSF for DistinguishedYoung Scholar (2022B1515020009), open research fund of Key Laboratory of Machine Intelligence and System Control, Ministry of Education (No. MISC-202407)

References

[1] 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

[2] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 2

[3] Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025. 2, 4

[4] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18000–18010. IEEE, 2023. 2

[5] chuan guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 7

[6] Manolo Canales Cuba and João Paulo Gois. Flowmotion: Target-predictive flow matching for realistic text-driven human motion generation. *arXiv preprint arXiv:2504.01338*, 2025. 2

[7] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023. 2

[8] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. 2024. 2

[9] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. Wandr: Intention-guided human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 927–936, 2024. 2

[10] John R Dormand and Peter J Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980. 3

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 2, 3, 5

[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*, pages 2021–2029, 2020. 6

[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 *CVPR*, pages 5152–5161, 2022. 1, 2, 3, 6, 7, 8

[14] 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 *ECCV*, pages 580–597. Springer, 2022. 2

[15] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 1900–1910. IEEE, 2024. 1, 2, 3, 5, 6, 7, 8

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 6

[17] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. *arXiv preprint arXiv:2503.13836*, 2025. 4, 6, 7

[18] Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M. Asano, Efstratios Gavves, Pascal Mettes, Björn Ommer, and Cees G. M. Snoek. Motion flow matching for human motion synthesis and editing. *CoRR*, abs/2312.08895, 2023. 2

[19] Guohong Huang, Ling-An Zeng, Zexin Zheng, Shengbo Gu, and Wei-Shi Zheng. Efficient explicit joint-level interaction modeling with mamba for text-guided HOI generation. In *IEEE International Conference on Multimedia and Expo, ICME 2025, Nantes, France, June 30 - July 4, 2025*, pages 1–6. IEEE, 2025. 2

[20] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 36:20067–20079, 2023. 2

[21] 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. 2024. 2

[22] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 4

[23] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 2

[24] Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun

Choi, Saim Shin, Jungho Kim, and Hyung Jin Chang. Personabooth: Personalized text-to-motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 22756–22765. Computer Vision Foundation / IEEE, 2025. 2

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[26] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816, 2023. 2

[27] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 5

[28] Heng Li, Xing Liufu, Xiaotong Lin, Jian Zhu, and Jian-Fang Hu. Efficient text-to-motion via multi-head generative masked modeling. In *IEEE International Conference on Multimedia and Expo, ICME 2025, Nantes, France, June 30 - July 4, 2025*, pages 1–6. IEEE, 2025. 2

[29] Yuan-Ming Li, Qize Yang, Nan Lei, Shenghao Fu, Ling-An Zeng, Jian-Fang Hu, Xihan Wei, and Wei-Shi Zheng. Irg-motionllm: Interleaving motion generation, assessment and refinement for text-to-motion generation. *arXiv preprint arXiv:2512.10730*, 2025. 2, 6, 7

[30] Zhuo Li, Mingshuang Luo, Ruibing Hou, Xin Zhao, Hao Liu, Hong Chang, Zimo Liu, and Chen Li. Morph: A motion-free physics optimization framework for human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14580–14589, 2025. 1

[31] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3

[32] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3

[33] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. Bridging the gap between human motion and action semantics via kinematic phrases. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[34] 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 *Forty-first International Conference on Machine Learning*.

[35] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 27872–27882. Computer Vision Foundation / IEEE, 2025. 2

[36] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6427–6436. IEEE, 2022. 1

[37] 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. 6

[38] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 27859–27871. Computer Vision Foundation / IEEE, 2025. 1

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4172–4182. IEEE, 2023. 2, 5, 7

[40] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: generating diverse human motions from textual descriptions. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022. 2, 6, 7

[41] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. BAMM: bidirectional autoregressive motion model. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XV*, pages 172–190. Springer, 2024. 1, 2, 7

[42] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: generative masked motion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 1546–1555. IEEE, 2024. 2, 5

[43] Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9955–9965, 2025. 2

[44] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 6, 7, 8

[45] 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, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 5

[46] Zeping Ren, Shaoli Huang, and Xiu Li. Realistic human motion generation with cross-diffusion models. 2024. 2

[47] Alessio Sampieri, Alessio Palma, Indro Spinelli, and Fabio Galasso. Length-aware motion synthesis via latent diffusion. 2024. 2

[48] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12026–12035. Computer Vision Foundation / IEEE, 2019. 3, 4, 6

[49] Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[50] Jiangxin Sun, Zihang Lin, Xintong Han, Jian-Fang Hu, Jia Xu, and Wei-Shi Zheng. Action-guided 3d human motion prediction. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021. Curran Associates Inc. 2

[51] Jiangxin Sun, Chunyu Wang, Huang Hu, Hanjiang Lai, Zhi Jin, and Jian-Fang Hu. You never stop dancing: Non-freezing dance generation via bank-constrained manifold projection. In *Advances in Neural Information Processing Systems*, 2022.

[52] Jianwei Tang, Jian-Fang Hu, Tianming Liang, Xiaotong Lin, Jiangxin Sun, Wei-Shi Zheng, and Jianhuang Lai. Human motion prediction via continual prior compensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2026. 2

[53] 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. 2

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 5, 7

[55] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. 2024. 2

[56] Runqi Wang, Caoyuan Ma, Guopeng Li, Hanrui Xu, Yuke Li, and Zheng Wang. You think, you act: The new task of arbitrary text to motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12012–12022, 2025. 1

[57] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 2

[58] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[59] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10086–10096, 2025. 2

[60] 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*. 2

[61] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4805–4815, 2018. 4, 5

[62] Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. Mogents: Motion generation based on spatial-temporal joint modeling. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1, 2, 4, 5, 6, 7, 8

[63] Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chain-hoi: Joint-based kinematic chain modeling for human-object interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 12358–12369. Computer Vision Foundation / IEEE, 2025. 2

[64] Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-t2m: A lightweight and fast model for text-to-motion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9797–9805, 2025. 2, 6, 7

[65] Ling-An Zeng, Gaojie Wu, Ancong Wu, Jian-Fang Hu, and Wei-Shi Zheng. Progressive human motion generation based on text and few motion frames. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2

[66] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Y ong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pages 14730–14740, 2023. 2, 6, 7

[67] Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 17592–17602. Computer Vision Foundation / IEEE, 2025. 6, 7

[68] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, pages 364–373, 2023. 7

[69] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *TPAMI*, 2024. 2, 6

[70] Zongye Zhang, Bohan Kong, Qingjie Liu, and Yunhong Wang. Towards robust and controllable text-to-motion via masked autoregressive diffusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 9326–9335, New

York, NY, USA, 2025. Association for Computing Machinery. 2

[71] 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

[72] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. 2024. 2

[73] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2024. 2

[74] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, pages 126–143. Springer, 2024. 2