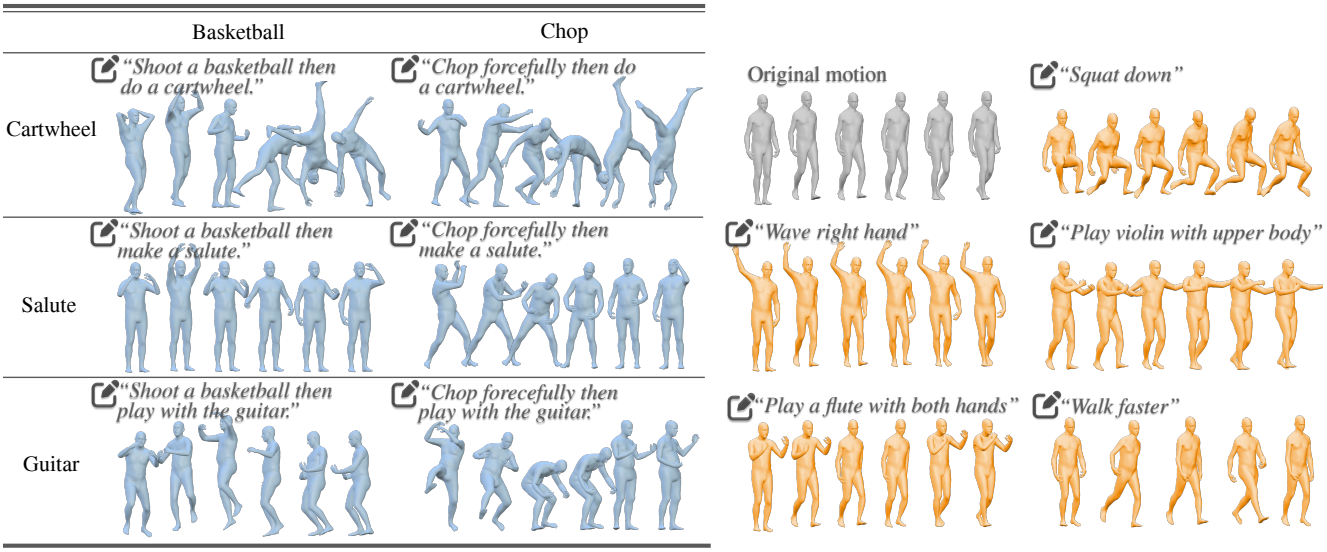
MotionMaster: Generalizable Text-Driven Motion Generation and Editing

Nan Jiang[1,2,3,7,8,9*] Yunhao Li[3,6,7,8*] Lexi Pang[1,3,5,7,8*]
Zimo He[4,2,7,8,9] Siyuan Huang[2,7 ✉] Yixin Zhu[3,1,7,8,9 ✉]

[1] Institute for AI, Peking University [2] Beijing Institute for General Artificial Intelligence (BIGAI)
[3] School of Psychological and Cognitive Sciences, Peking University [4] School of Computer Science, Peking University
[5] Yuanpei College, Peking University [6] School of Foreign Languages, Peking University
[7] State Key Lab of General AI [8] Beijing Key Laboratory of Behavior and Mental Health, Peking University
[9] Embodied Intelligence Lab, PKU-Wuhan Institute for Artificial Intelligence
[*] Equal contribution ✉ yixin.zhu@pku.edu.cn, syhuang@bigai.ai https://jnnan.github.io/motionmaster



(a) End-to-end sequential text-to-motion generation (b) Text-guided motion editing

Figure 1. **The MotionMaster framework.** By finetuning a pretrained MLLM on large-scale motion data, MotionMaster enables unified text-guided human motion generation and editing within an end-to-end framework. (a) MotionMaster generates compositional action sequences from textual instructions, successfully producing action combinations never seen during training (*e.g.*, "shoot a basketball then do a cartwheel") by leveraging action semantics inherited from the pretrained MLLM. (b) Given an original motion, MotionMaster applies precise text-guided edits to specific body parts or motion properties while preserving the remainder of the sequence, demonstrating precision and fine-grained control.

Abstract

Synthesizing realistic human motion from natural language holds transformative potential for animation, robotics, and virtual reality. Recent methods handle single-action sequences and simple textual instructions, yet multi-action compositions and precise editing remain elusive due to limited data diversity, inadequate representations, and fragmented pipelines. Critically, most existing methods train motion generation models from scratch, failing to exploit the rich action semantics and long-horizon reasoning already encoded in pretrained Multimodal Large Language Models (MLLMs). Here we show that finetuning a pretrained MLLM with large-scale motion data yields strong zero-shot generalization across diverse text-guided motion generation and editing tasks. We present MotionMaster, a unified framework built on three components: MotionGB, a 10,000-hour dataset expanded from 400 hours of verified motion capture via spatial-temporal augmentation; an FSQ-based tokenizer that preserves both local joint accuracy and global trajectory coherence; and a finetuned MLLM with motion and language tokens in a shared embedding space. MotionMaster outperforms prior methods by 41.6% in multi-action semantic consistency and 20.8% in body-part composition. These results demonstrate that pretraining knowledge from MLLMs transfers effectively to motion understanding, opening a viable path toward general-purpose motion intelligence.

1. Introduction

Text-driven human motion generation has emerged as a fundamental challenge in computer graphics and computer vision, with applications spanning animation [10, 18, 58], virtual reality [28, 55], and humanoid robot training [20, 21]. While recent advances have shown promising results in generating simple, single-action motions from text descriptions [11, 18], existing methods struggle with real-world scenarios that demand complex multi-action sequences or precise motion editing capabilities.

Current approaches suffer from three compounding limitations. First, the data foundation is inadequate—most methods [17, 18, 43] rely on small datasets with coarse annotations that fail to capture motion nuances and semantic diversity. Even methods trained on larger datasets [14, 19, 39, 61] struggle to generalize to complex text instructions, particularly those requiring compositional understanding. Second, existing motion representations [14, 46] fail to balance local joint accuracy with global trajectory coherence, a problem that intensifies during large-scale training. Third, multi-action generation [45] and body-part editing [46] rely on fragmented pipelines that compose separately generated components, revealing a lack of genuine end-to-end understanding of unified motion semantics. Most fundamentally, existing methods [14, 29, 39] train motion models from scratch, forgoing the rich action semantics and long-horizon reasoning already encoded in pretrained MLLMs—a limitation that collectively prevents generalization across complex language-motion dependencies and fine-grained specifications.

To address these challenges, we introduce MotionMaster, a unified end-to-end framework that finetunes a pretrained MLLM on large-scale motion data for generalizable text-guided motion generation and editing. Our framework rests on three contributions.

MotionGB We construct a motion dataset containing 10,000 hours of richly annotated data. Starting from 400 hours of manually verified motion capture data, we generate verbal descriptions of pose configurations at each timestep, followed by MLLM summarization to produce multi-level motion descriptions. Through systematic augmentation—temporal concatenation, body-part composition, and motion editing operations—we expand this foundation while maintaining precise motion-text correspondence, providing the semantic breadth necessary for understanding complex motion patterns.

Motion tokenization We develop a SMPL-X-based motion discretization method that preserves both local joint accuracy and global trajectory coherence. Frame-by-frame local keypoint features, including yaw angular speed and joint positions, are encoded into discrete tokens via Finite Scalar Quantization (FSQ), while a reconstruction loss evaluated on full sequences in global coordinates prevents trajectory drift. This localized representation naturally maximizes codebook utilization without requiring unique codes for every global pose configuration.

Unified motion-language modeling By adapting the Qwen2.5-VL architecture with motion and language tokens sharing the same embedding space, MotionMaster enables direct cross-modal fusion within a single autoregressive model. The pretrained MLLM contributes prior knowledge of action semantics and long-horizon dependencies that are difficult to acquire through post-hoc composition. To ensure exposure to the full spectrum of motion semantics, we further introduce semantic balancing during finetuning, adjusting sampling probabilities based on semantic density to prevent overfitting to frequently occurring patterns.

We evaluate MotionMaster using an automated assessment framework that renders generated motions into videos and uses Gemini [56] as the evaluator, validated with a correlation coefficient of 0.89 against human judgments. MotionMaster outperforms existing methods by 26.8% on out-of-distribution (OOD) single motion generation, by 41.6% in semantic consistency for multi-action temporal composition, and by 20.8% in accuracy for spatial body-part composition. Beyond quantitative gains, MotionMaster exhibits emergent capabilities—including zero-shot motion style transfer, physics-aware motion correction, and compositional reasoning over complex action sequences—that arise from genuine semantic understanding inherited from the pretrained MLLM rather than memorized patterns.

Our key contributions are summarized as:

- **MotionGB: A 10,000-hour richly annotated motion-language dataset**—We provide the semantic breadth and quality necessary for understanding complex motion patterns, from simple actions to intricate multi-part sequences, supporting both generation and editing tasks.
- **A scalable SMPL-X-based motion tokenization method**—We preserve both local joint accuracy and global trajectory coherence, achieving state-of-the-art (SOTA) motion fidelity under large-scale training.
- **MotionMaster: A unified end-to-end framework for text-guided motion generation and editing**—By finetuning a pretrained MLLM with large-scale motion data, we demonstrate strong zero-shot generalization across diverse text-motion tasks.

2. Related Work

Motion generation Motion generation has advanced significantly through recent research, evolving from early approaches using conditional Variational Autoencoders (VAEs) [3, 16, 23, 42, 43], Generative Adversarial Networks (GANs) [37, 62], contrastive learning [35, 44, 57], and masked modeling [18, 46] to generate motion from simple text descriptions. Works leveraging diffusion models [9, 12, 28, 30, 31, 34, 58, 67, 69] marked a break-

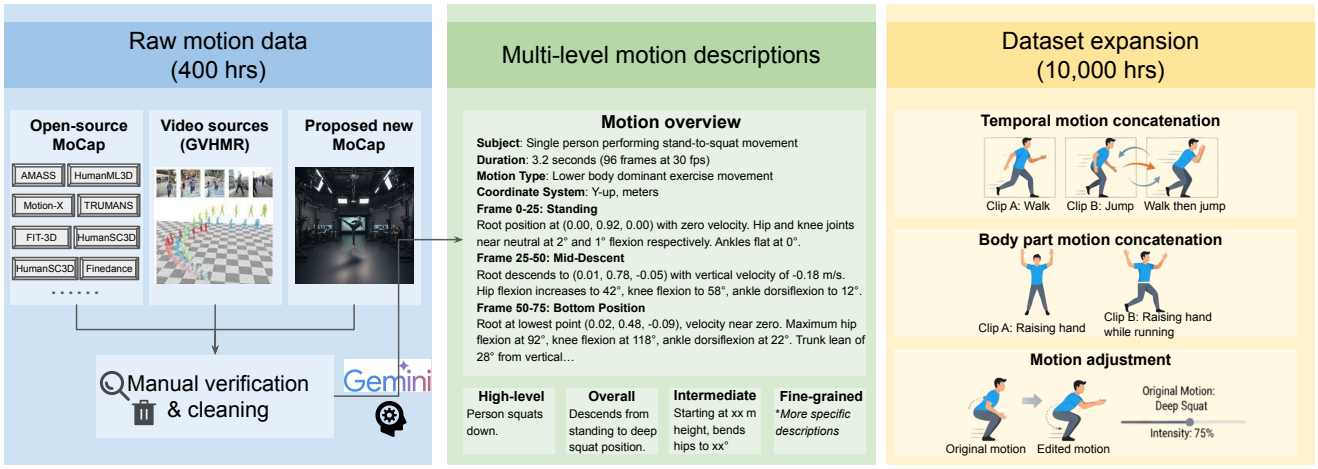


Figure 2. **Overview of MotionGB construction pipeline.** (a) 400 hours of raw motion data are collected from open-source motion capture databases, video sources, and proprietary recordings. (b) After manual verification and cleaning, each sequence is annotated with multi-level motion descriptions generated via Gemini [56]. (c) The dataset is expanded to 10,000 hours through three augmentation strategies: temporal concatenation (combining sequential actions), body-part concatenation (merging movements from different body parts), and fine-grained motion adjustment (applying parametric modifications to create editing pairs).

through, with MDM [58] and MotionDiffuse [69] introducing diffusion processes for human motion, and MLD [9] improving computational efficiency through latent space operations. MotionLab [19] leverages rectified flow for unified motion generation and editing. However, these methods predominantly rely on CLIP as their text encoder, limiting their ability to understand complex motion-specific language and temporal relationships. In parallel, MLLM-based approaches [8, 14, 27, 39, 63, 66, 70, 71] have emerged. T2M-GPT [66] and MotionGPT [27] treat motion as discrete tokens for autoregressive generation, while MotionLLM [63] and ScaMo [39] incorporate human motion as an additional modality alongside text. MotionStreamer [64] proposes a diffusion-based autoregressive model for temporal coherence. Despite these advances, a fundamental limitation persists: most of these methods train from scratch, forgoing the rich action semantics already encoded in pre-trained MLLMs, and consequently fail to establish genuine semantic correspondence between linguistic descriptions and motion expressions—particularly for complex scenarios involving multiple actions, limb nuances, or context-dependent behaviors.

Motion editing Initial motion editing work focused on trajectory manipulation [15, 32, 38], skeletal retargeting [1], and emotion modulation [59], which evolved into motion style transfer techniques [2, 6, 13, 49, 54, 65]. Semantic motion editing has since developed along several paradigms. Early embedding-based methods [22, 57] encoded semantic information in latent representations but lack fine-grained disentanglement. Diffusion models enabled more sophisticated editing through inpainting [31, 46, 58, 69] and compositional techniques—including temporal composition [3, 51], spatial body-part control [4, 43], and timeline-based manipulation [45]. More recent approaches leverage foundation models, with FineMoGen [68] and COMO [24] demonstrating strong performance on annotated datasets. To handle arbitrary unannotated inputs, newer methods based on diffusion models [7, 29, 33] and flow matching [19] directly condition on both source motion and textual instructions. Despite these advances, existing editing methods remain tightly coupled to their training distributions, exhibiting limited generalization when handling arbitrary semantic textual descriptions.

Motion datasets Human motion datasets have established critical foundations for multimodal motion generation and editing. Early motion-only datasets [26, 52] provided diverse capture data across multiple modalities and action categories, while motion-captured datasets [16, 25, 40] offered extensive repositories of 3D skeletal sequences. The integration of textual annotations marked a significant evolution: KIT-ML [47] introduced structured motion-language pairs, and BABEL [48] extended AMASS [40] with body-level annotations. HumanML3D [17] further enriched this paradigm by associating each motion with multiple text descriptions. Recent efforts to recover motion from video [36, 60] have scaled up data volume, though quality often suffers from camera movement and occlusions. Even larger collections [14, 34, 39] have been proposed, yet the field still lacks the emergent generalization seen in language and vision models—a gap that motivates MotionGB's emphasis on richer linguistic annotations, multi-level descriptions, and diverse editing examples.

3. The MotionGB Dataset

We present MotionGB, a large-scale motion dataset of richly annotated human movements designed for training

generative motion models. As Fig. 2 shows, our construction pipeline comprises multi-level semantic annotation and systematic data augmentation. **??** provides additional statistical analysis.

3.1. Raw Motion Data Collection

MotionGB derives from 400 hours of raw motion data from three sources. First, we incorporate open-source motion capture datasets that provide high-fidelity skeletal movements. Second, we extract motion data from video sources using GVHMR [53], a SOTA method for recovering SMPL-X body model representations from video. Third, we include proprietary motion recordings that capture expressive, nuanced human movements that are underrepresented in existing datasets. To preserve semantic integrity, we maintain complete motion sequences in their original form, except for long dance performances. Each sequence is manually validated by trained annotators, who remove motions exhibiting severe artifacts, such as foot sliding, joint jitter, or anatomically impossible poses.

3.2. Multi-Level Motion Description

Our annotation pipeline transforms raw motion into multi-level textual descriptions through two stages. First, we extract frame-wise motion reports capturing joint angles, limb positions, torso orientation, and movement velocities. These quantitative reports then feed into Gemini [56], which generates descriptions at four semantic levels: high-level intent ("morning exercise"), overall action ("walks forward while stretching arms"), intermediate phases ("lifts right arm to shoulder height, then extends upward"), and fine-grained details ("bends left knee to 45 degrees while rotating right wrist clockwise"). This hierarchy enables left-right distinction, precise degree specification, and seamless integration between high-level semantics and low-level kinematics.

3.3. Dataset Expansion Strategy

We employ three augmentation pipelines to expand MotionGB from 400 to 10,000 hours while simultaneously generating motion-editing pairs for training.

Temporal combination We concatenate 2–3 motion sequences to create multi-action sequences, with descriptions grammatically modified to reflect the combined actions. To ensure natural transitions, we train a motion in-betweening model on raw sequences to generate 1-second transition segments. **??** provides detailed validation of our in-betweening model's quality, including quantitative metrics and a perceptual user study.

Body-part concatenation We blend movements from different body parts to produce concurrent actions, such as "jogging while typing." For each raw motion, we randomly select 10 other motions and systematically replace specific body-part movements, restricting lower-body source candidates to sequences with active lower-body engagement. This generates abundant training pairs for body-part-specific editing tasks, enabling capabilities such as "replace lower body with running."

Fine-grained motion adjustment We apply 24 types of parametric modifications (*e.g.*, joint transforms, rotations, speed variations, and styles). Each modification produces a new motion, accompanied by a precise description of the applied change (*e.g.*, "raise right arm higher," "speed up the motion"), thereby creating a rich set of instruction-following examples for fine-grained motion control.

4. Motion Tokenization

Our motion tokenization consists of three components: (i) extracting localized motion features from raw joint positions, (ii) encoding and quantizing these features via FSQ, and (iii) reconstructing motion with global supervision.

4.1. Finite Scalar Quantization

We employ FSQ to discretize continuous motion sequences into discrete tokens. FSQ offers stable training and predictable codebook utilization by quantizing each latent dimension independently.

Given localized motion features (detailed in Sec. 4.2), we encode them through a 1D convolutional encoder E:

$$\mathbf{z} = E(\mathbf{f}) \in \mathbb{R}^{T' \times D}, \tag{1}$$

where $\mathbf{f} \in \mathbb{R}^{T \times 85}$ represents the input features for T frames, $T' < T$ is the compressed temporal dimension, and D is the latent dimension. The encoder consists of 4 convolutional layers with kernel size 4 and stride 2, progressively reducing temporal resolution while learning abstract motion representations.

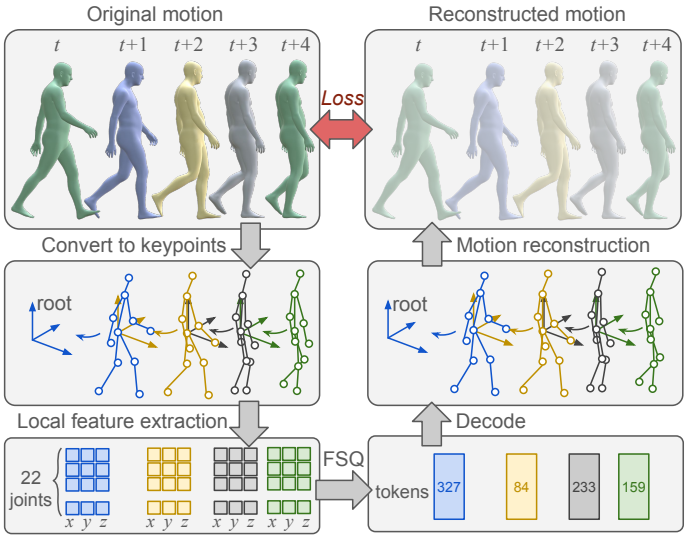
FSQ quantizes each latent element independently to discrete levels:

$$\hat{z}_{i,d} = \text{round}(z_{i,d} \cdot L_d)/L_d, \tag{2}$$

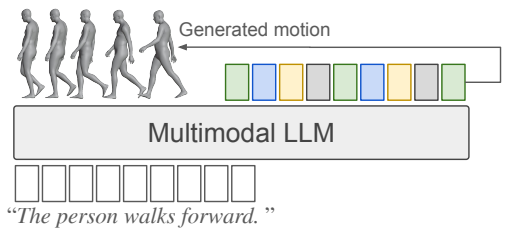
where L_d defines the number of quantization levels for dimension d. This element-wise quantization ensures full codebook utilization, as each dimension can independently take any of its allowed values. During training, we use straight-through estimation to allow gradient flow through the quantization operation. The quantized latents are decoded through a symmetric decoder that uses transposed 1D convolutions to upsample the compressed representation back to the original temporal resolution.

4.2. Localized Feature Extraction

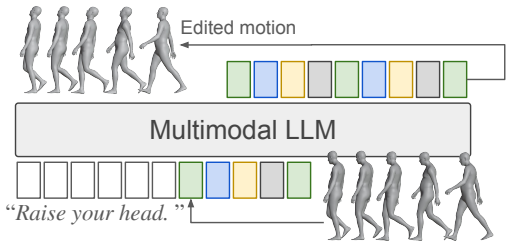
Given a motion sequence with joint positions $\mathbf{p}_t \in \mathbb{R}^{J \times 3}$ for J joints and root orientation $\mathbf{r}_t$ at time t, we extract an 85-dimensional feature vector representing the transition from frame t to frame $t+1$ in a localized coordinate system.



(a) Motion representation



(b) Text-to-motion generation



(c) Text-guided motion editing

Figure 3. **Overview of MotionMaster.** (a) The FSQ-based motion tokenizer encodes joint positions into localized features, quantizes them into discrete tokens, and supervises reconstruction via a loss computed in global coordinates. (b) For text-to-motion generation, the finetuned MLLM autoregressively decodes motion tokens conditioned on a text prompt. (c) For text-guided editing, the original motion is provided as additional context, and the MLLM selectively modifies the relevant tokens while preserving the remainder of the sequence.

Step 1: Extract yaw angle We compute the horizontal orientation by applying the root rotation to the forward vector and extracting the yaw component:

$$\theta_t = \text{atan2}(\mathbf{r}_t \cdot [0, 0, 1]). \tag{3}$$

Step 2: Compute orientation change The frame-to-frame yaw difference captures turning motion:

$$\Delta\theta_t = \theta_{t+1} - \theta_t. \tag{4}$$

Step 3: Transform to relative coordinates We express all joints relative to the previous frame's coordinate system. The previous root is projected onto the ground plane under a y-up convention:

$$\mathbf{p}_t^{\text{root}} = [\mathbf{p}_{t,0,x},\ 0,\ \mathbf{p}_{t,0,z}]. \tag{5}$$

The relative positions are then rotated by the negative of the previous frame's yaw:

$$\mathbf{p}'_{t+1} = R_{-\theta_t}(\mathbf{p}_{t+1} - \mathbf{p}_t^{\text{root}}), \tag{6}$$

where $R_{-\theta_t}$ applies a 2D rotation to the horizontal components while preserving height:

$$R_{-\theta_t} = \begin{bmatrix} \cos(-\theta_t) & 0 & -\sin(-\theta_t) \\ 0 & 1 & 0 \\ \sin(-\theta_t) & 0 & \cos(-\theta_t) \end{bmatrix}. \tag{7}$$

Step 4: Construct feature vector The final feature concatenates the yaw change with flattened relative joint positions:

$$\mathbf{f}_t = [\Delta\theta_t,\ \text{flatten}(\mathbf{p}'_{t+1})] \in \mathbb{R}^{85}. \tag{8}$$

This localized representation ensures that similar motion patterns (*e.g.*, walking, turning) produce similar features regardless of their absolute position in world space, maximizing codebook reuse while preserving reconstruction fidelity.

4.3. Global Reconstruction and Loss Functions

While tokenization operates on localized features to improve efficiency, we recover SMPL-X parameters via gradient descent using a joint-to-joint loss. The reconstruction process inverts the feature extraction process, but quantization errors can accumulate during this sequential inversion. We address this by explicitly supervising global positions.

Given reconstructed features $\hat{\mathbf{f}}_t = [\Delta\hat{\theta}_t, \hat{\mathbf{p}}'_{t+1}]$, we recover global motion iteratively.

Step 1: Accumulate orientation Starting from an initial pose, we integrate yaw changes:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \Delta\hat{\theta}_t. \tag{9}$$

Step 2: Transform to world space We apply the accumulated rotation and translation:

$$\hat{\mathbf{p}}_{t+1} = R_{\hat{\theta}_t}(\hat{\mathbf{p}}'_{t+1}) + \hat{\mathbf{p}}_t^{\text{root}}. \tag{10}$$

Since errors in $\Delta\hat{\theta}_t$ and relative positions accumulate over time, we supervise global joint positions and velocities directly to prevent drift:

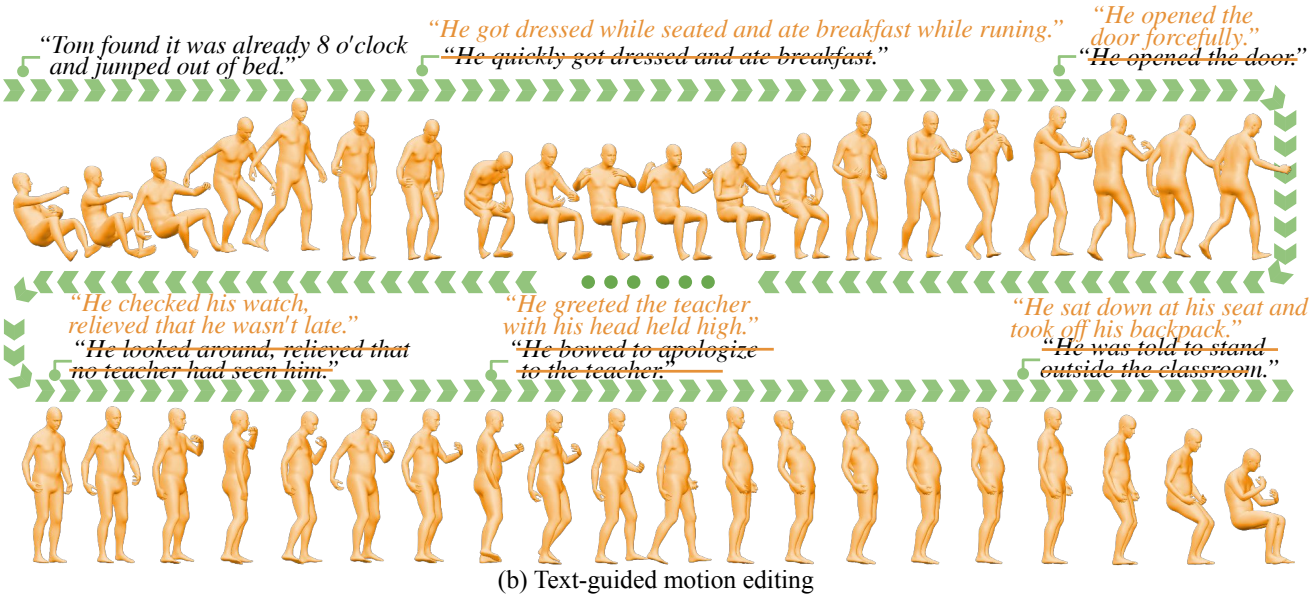
$$\mathcal{L}_{\text{global}} = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \|\mathbf{p}_{t,j} - \hat{\mathbf{p}}_{t,j}\|_2^2, \tag{11}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{(T-1)J} \sum_{t=1}^{T-1} \sum_{j=1}^{J} \|(\mathbf{p}_{t+1,j} - \mathbf{p}_{t,j}) - (\hat{\mathbf{p}}_{t+1,j} - \hat{\mathbf{p}}_{t,j})\|_2^2. \tag{12}$$

4.4. Coarse-to-Fine IK Solver

Converting generated keypoint trajectories into SMPL-X parameters requires recovering local joint rotations from

(a) Text-to-motion generation



(b) Text-guided motion editing

Figure 4. **Demonstration of MotionMaster's multi-action generation and editing capabilities on a long narrative sequence.** (a) MotionMaster autoregressively generates motion segments from sequential text descriptions, producing a coherent long-horizon motion sequence. (b) Text-guided edits are then applied to the previously generated motion, modifying specific actions or body-part movements while preserving the overall narrative context.

sparse 3D joint positions—an inherently under-constrained Inverse Kinematics (IK) problem in which errors accumulate rapidly along the kinematic chain. Directly minimizing positional error with respect to joint rotations often yields physically implausible configurations, such as severely twisted wrists or ankles.

We address this with a two-stage coarse-to-fine IK solver. In the first stage, we optimize within the latent manifold of VPoser [41], a learned generative prior of biomechanically plausible human poses. By optimizing the low-dimensional latent embedding rather than explicit joint rotations, the search space is constrained to anatomically valid configurations. In the second stage, we use the solution from the first stage as a robust initialization and fine-tune joint

rotations to strictly regress the target joint positions. This coarse-to-fine formulation achieves high reconstruction fidelity while ensuring physical plausibility throughout. **??** details our tokenizer architecture and the SMPL-X joint-to-parameter conversion pipeline.

5. The MotionMaster

We detail the core components of MotionMaster: our unified motion-language modeling approach and the semantic balancing technique that addresses dataset imbalance.

5.1. Unified Motion-Language Modeling

We extend Qwen2.5-VL to process both text and motion tokens in a shared embedding space. Using the motion tokenization described in Sec. 4, we integrate discrete motion codes directly into the MLLM's existing vocabulary. Rather than expanding the vocabulary, we substitute the least-used text tokens—typically special characters and obsolete symbols—in Qwen2.5-VL's original codebook with our motion tokens. We further introduce special tokens <SOM> (start of motion) and <EOM> (end of motion) to demarcate motion sequences within mixed-modality inputs.

Training strategy During finetuning, we freeze embeddings of active text tokens to preserve linguistic knowledge, training only the repurposed motion token embeddings and transformer weights. We use causal attention for autoregressive generation: $P(m_t \mid t_{\text{prompt}}, m_{<t})$.

Position encoding Our Rotary Position Embeddings (RoPE) implementation uses distinct counters for text and motion modalities. A global counter tracks position indices for all text tokens sequentially. Upon encountering a <SOM> token, a separate motion-specific counter is initialized to zero and increments independently for that sequence. This ensures each modality operates within its own positional context even when interleaved, avoiding cross-modal position interference.

5.2. Semantic Balancing

Raw motion datasets contain severe semantic imbalances—walking and standing motions vastly outnumber complex actions such as dancing—which cause models to overfit to common patterns. Our semantic balancing ensures uniform coverage of the motion space during training.

Density estimation For each motion-text pair (m_i, t_i), we compute a semantic embedding $e_i = \phi(t_i)$ using a T5 text encoder [50]. Local density is estimated via a Gaussian kernel:

$$\rho_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \exp\left(-\frac{\|e_i - e_j\|^2}{2\sigma^2}\right), \tag{13}$$

where $\mathcal{N}_k(i)$ denotes the k-nearest neighbors of e_i in semantic space, σ controls the kernel bandwidth, and α adjusts rebalancing strength. Sampling probabilities are assigned as $p_i \propto \rho_i^{-\alpha}$.

Weighted sampling During training, we sample batches according to p_i rather than uniformly. Motions in semantically sparse regions are sampled more frequently, while overrepresented patterns receive proportionally less exposure, ensuring robust learning across the full spectrum of motion semantics.

6. Experiments

We conduct comprehensive experiments to evaluate MotionMaster across multiple dimensions: motion tokenizer quality (Sec. 6.1), generalization on diverse generation and editing tasks (Sec. 6.2), and ablation studies on data scale, model size, and joint training (Sec. 6.3).

6.1. Motion Tokenizer Evaluation

Dataset and training setup We evaluate on MotionGB, partitioned into 90% for training (MotionGB-train) and 10% for evaluation (MotionGB-test). All methods, including baselines, are trained on MotionGB-train to ensure fair comparison.

Baseline methods We compare our SMPL-X-based motion discretization against four SOTA motion representation methods: T2M-GPT [66], MoMask [18], MMM [46], and MotionMillion [14].

Evaluation metrics We evaluate along two dimensions. **(i) Local motion accuracy:** rotational error of human joints in SMPL-X representation (average angular deviation in degrees) and localized joint position error (mean Euclidean distance in the local coordinate frame). **(ii) Global trajectory coherence:** rotational error of the root in global coordinates, global joint position error, and joint velocity error capturing temporal consistency.

Results As Tab. 3 shows, our tokenizer achieves superior joint position accuracy across both local and global metrics, demonstrating effective balance between fine-grained joint accuracy and trajectory coherence. The relatively large rotation error is attributable to the undefined rotation along each bone's axial direction, which lacks direct restriction. Ablation studies justifying our loss function design are provided in **??**.

6.2. Motion Generation and Editing Evaluation

Datasets We construct MotionGB-test-lite by randomly sampling 400 motions from MotionGB-test, covering (i) single motion sequences, (ii) multi-action sequences, (iii) concurrent actions, and (iv) adjusted motions. For editing evaluation, categories (iii) and (iv) are structured as triples {original motion, edited motion, edit text}.

Baseline methods We evaluate MotionMaster against SOTA methods including MoMask [18], MMM [46], T2M-GPT [66], MotionLab [19], MotionFix [5], and MotionMillion [14]. Methods marked † are retrained on MotionGB for both motion representation and generation.

Table 1. **Quantitative comparison of MotionMaster against SOTA baselines and ablation studies.** MotionMaster achieves superior performance across most metrics for both motion generation and editing. The relatively lower diversity scores reflect a known trade-off between diversity and semantic fidelity rather than a limitation of the approach.

Method	Single Motion Generation						Long Sequence Generation			Motion Editing					
	Semantic↑	Diversity↑	Phys.↑	R-Precision↑			Semantic↑	Diversity↑	Phys.↑	Semantic↑	Diversity↑	Phys.↑	R-Precision↑		
				R@1	R@2	R@3							R@1	R@2	R@3
T2M-GPT† [66]	6.72	1.72	6.40	0.15	0.24	0.29	2.66	1.89	5.65	-	-	-	-	-	-
MotionM. [14]	6.98	**3.10**	6.70	0.09	0.14	0.18	2.92	**3.88**	**7.40**	-	-	-	-	-	-
MMM† [46]	7.20	0.86	6.70	0.16	0.22	0.30	3.34	1.03	4.25	4.10	**1.55**	5.35	0.14	0.22	0.30
MoMask† [18]	6.68	1.51	5.80	0.11	0.15	0.20	2.36	1.80	4.45	5.52	1.47	5.60	0.19	0.28	0.36
MotionLab [19]	6.90	2.10	5.65	0.13	0.21	0.26	3.02	2.83	4.35	5.60	0.16	7.35	0.21	0.27	0.34
Motionfix† [5]	-	-	-	-	-	-	-	-	-	7.02	1.12	7.35	0.27	0.47	0.56
Ours	**9.88**	1.62	**8.75**	0.24	0.33	0.39	**7.50**	2.47	7.05	**9.10**	0.50	**8.65**	**0.77**	**0.92**	**0.93**
Ours 3B model	8.78	1.60	7.70	**0.31**	**0.43**	**0.46**	5.35	2.52	5.60	8.77	0.41	8.25	0.75	0.89	0.90
Ours 50% data	7.90	1.76	7.55	0.24	0.36	0.42	5.52	2.52	5.40	8.75	0.59	8.60	0.66	0.88	0.91
Ours w/o semantic	8.58	1.41	7.80	0.27	0.40	0.44	-	-	-	-	-	-	-	-	-

Table 2. **Mutual benefits of joint training on generation and editing.** Joint training consistently outperforms task-specific training on both generation and editing, confirming that the two tasks are complementary and mutually reinforcing.

Method	Single Motion Generation						Long Sequence Generation			Motion Editing					
	Semantic↑	Diversity↑	Phys.↑	R-Precision↑			Semantic↑	Diversity↑	Phys.↑	Semantic↑	Diversity↑	Phys.↑	R-Precision↑		
				R@1	R@2	R@3							R@1	R@2	R@3
Generation Only	7.40	**2.19**	7.40	0.19	0.28	0.33	3.20	**3.05**	4.90	-	-	-	-	-	-
Editing Only	-	-	-	-	-	-	-	-	-	8.30	0.49	8.40	**0.78**	0.89	**0.94**
Joint Training	**9.88**	1.62	**8.75**	**0.24**	**0.33**	**0.39**	**7.50**	2.47	**7.05**	**9.10**	**0.50**	**8.65**	0.77	**0.92**	0.93

Table 3. **Motion tokenizer evaluation on MotionGB-test.** Our tokenizer achieves the best joint position accuracy across both local and global metrics. Lower values indicate better performance.

Method	Local Accuracy		Global Coherence		
	Rot. (°)	Pos. (cm)	Rot. (°)	Pos. (cm)	Vel. (cm/s)
T2M-GPT [66]	5.40	11.92	11.89	16.92	20.1
MoMask [18]	**3.77**	9.56	13.46	15.74	19.8
MotionM. [14]	4.28	12.31	12.68	21.96	20.6
MMM [46]	19.41	10.50	18.53	18.84	29.5
Ours	7.55	**9.14**	**10.13**	**9.53**	**15.3**

Evaluation metrics We introduce an evaluation framework combining semantic alignment with standard motion quality metrics. **Semantic alignment:** we render generated motions into video and query Gemini [56] to assess whether the actions match the textual description; for editing tasks, original and edited motions are rendered side-by-side and scored for correctness and conciseness on a 0–10 scale, validated at 0.89 correlation with human judgments (see **??**). **Physical plausibility:** Gemini assesses the physical realism of generated motions following the same protocol. **Retrieval precision** and **diversity** are computed following Guo et al. [17].

Results Qualitative results are shown in Fig. 4; quantitative results are presented in Tab. 1. MotionMaster achieves the highest performance across most metrics on both generation and editing tasks. The relatively lower diversity scores reflect a known trade-off between diversity and semantic fidelity. A human preference study further validates these findings; details are provided in **??**.

6.3. Ablation Studies

We examine two aspects. First, Tab. 1 shows that both larger datasets and increased model capacity consistently improve performance, and that semantic balancing yields significant gains—confirming the importance of all three design choices. Additional scaling experiments are provided in **??**. Second, Tab. 2 demonstrates that joint training on generation and editing provides significant advantages for both tasks, confirming their complementary nature.

7. Conclusion

We present MotionMaster, an end-to-end framework for text-guided human motion generation and editing. By fine-tuning a pretrained MLLM on MotionGB—a 10,000-hour richly annotated dataset—and equipping it with an FSQ-based motion tokenizer and semantic balancing, MotionMaster achieves strong zero-shot generalization across diverse motion tasks, from complex multi-action sequences to precise body-part editing.

Two key insights emerge from our results: dataset scale and diversity are critical for motion generation capabilities, and jointly training generation and editing creates synergistic benefits where each task reinforces the other. Beyond these, MotionMaster exhibits emergent zero-shot capabilities never explicitly supervised, suggesting that MLLM pre-training transfers more broadly to motion understanding than previously assumed. Together, these findings suggest that large-scale multimodal pretraining is a promising path toward general-purpose human motion intelligence.

Acknowledgement This work is supported in part by the Brain Science and Brain-like Intelligence Technology—National Science and Technology Major Project (2025ZD0219400), the National Natural Science Foundation of China (62376009), the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone. This project would not have been possible without the year-long collaboration with Virtual Point, whose partnership has been instrumental to this work.

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):1–14, 2020. 3

[2] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):1–11, 2020. 3

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

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

[5] Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3d human motion editing. In *SIGGRAPH Asia Conference Proceedings*, 2024. 7, 8

[6] Ziyi Chang, Edmund J. C. Findlay, Haozheng Zhang, and Hubert P. H. Shum. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. In *International Conference on Computer Graphics Theory and Applications (GRAPP)*, 2023. 3

[7] Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms. *arXiv preprint arXiv:2410.18977*, 2024. 3

[8] 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. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1—–15, 2025. 3

[9] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[10] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[11] Jieming Cui, Tengyu Liu, Ziyu Meng, Jiale Yu, Ran Song, Wei Zhang, Yixin Zhu, and Siyuan Huang. Grove: A generalized reward for learning open-vocabulary physical skill. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

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

[13] Yuzhu Dong, Andreas Aristidou, Ariel Shamir, Moshe Mahler, and Eakta Jain. Adult2child: Motion style transfer using cyclegans. In *ACM SIGGRAPH Motion, Interaction and Games (MIG)*, 2020. 3

[14] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2025. 2, 3, 7, 8

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

[16] 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*, 2020. 2, 3

[17] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 7, 8

[19] Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2025. 2, 3, 7, 8

[20] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning (CoRL)*, 2024. 2

[21] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024. 2

[22] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 3

[23] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 2

[24] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2024. 3

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014. 3

[26] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *International Conference on Multimedia*, 2018. 3

[27] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[28] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[29] Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3

[30] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 2

[31] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 3

[32] Manmyung Kim, Kyunglyul Hyun, Jongmin Kim, and Jehee Lee. Synchronized multi-character motion editing. *ACM Transactions on Graphics (TOG)*, 28(3):1–9, 2009. 3

[33] Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. Simmotionedit: Text-based human motion editing with motion similarity prediction. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[34] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[35] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[36] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

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

[38] Noah Lockwood and Karan Singh. Biomechanically-inspired motion path editing. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2011. 3

[39] 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 *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3

[40] 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 International Conference on Computer Vision (ICCV)*, 2019. 3

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[42] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 2

[43] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[44] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 2

[45] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2024. 2, 3

[46] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 7, 8

[47] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 4(4):236–252, 2016. 3

[48] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[49] Ziyun Qian, Zeyu Xiao, Xingliang Jin, Dingkang Yang, Mingcheng Li, Zhenyi Wu, Dongliang Kou, Peng Zhai, and Lihua Zhang. Umsd: High realism motion style transfer via unified mamba-based diffusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. 3

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 7

[51] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 3

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

[53] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024. 4

[54] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[55] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 3, 4, 8

[57] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[58] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[59] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. In *SIGGRAPH Conference Proceedings*, 1995. 3

[60] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Qin Jin, and Zongqing Lu. Quo vadis, motion generation? from large language models to large motion models. *arXiv preprint arXiv:2410.03311*, 2024. 3

[61] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu. Scaling large motion models with million-level human motions. In *Proceedings of International Conference on Machine Learning (ICML)*, 2025. 2

[62] 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 AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[63] Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. Motionllm: Multimodal motion-language learning with large language models. *arXiv preprint arXiv:2405.20340*, 2024. 3

[64] 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 International Conference on Computer Vision (ICCV)*, 2025. 3

[65] Wenjie Yin, Yi Yu, Hang Yin, Danica Kragic, and Mårten Björkman. Scalable motion style transfer with constrained diffusion generation. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 3

[66] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 7, 8

[67] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 2

[68] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[69] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(6):4115–4128, 2024. 2, 3

[70] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2024. 3

[71] 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. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 3