

Go to Zero: Towards Zero-shot Motion Generation with Million-scale Data

Ke Fan
Shanghai Jiao Tong University

Shunlin Lu
CUHK, Shenzhen

Minyue Dai
Fudan University

Runyi Yu
HKUST

Lixing Xiao
Zhejiang University

Zhiyang Dou
HKU

Junting Dong
Shanghai AI Laboratory

Lizhuang Ma[†]
Shanghai Jiao Tong University, East China Normal University

Jingbo Wang
Shanghai AI Laboratory

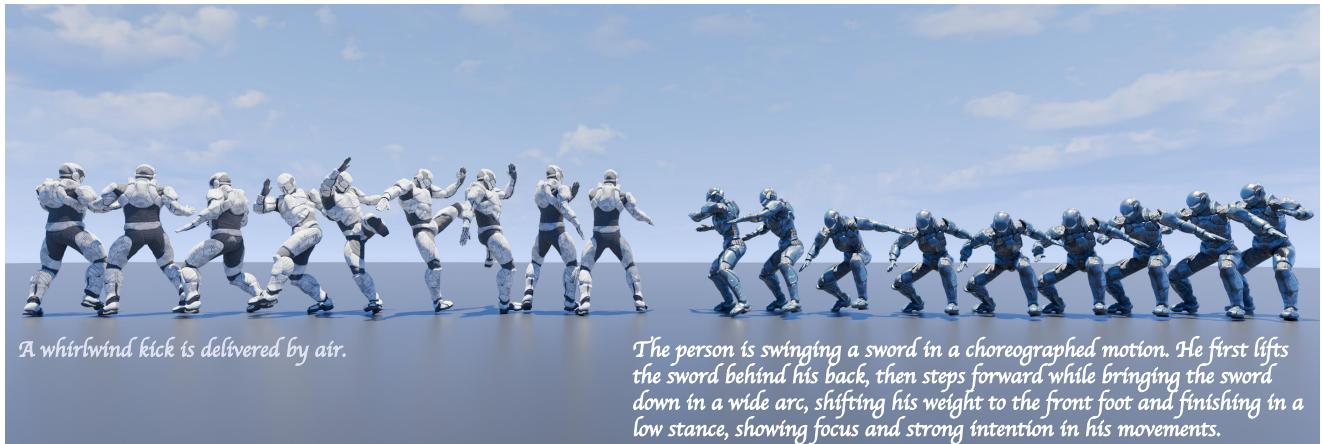


Figure 1. We present Go to Zero, where we can deal with out-domain and complex compositional motions.

Abstract

Generating diverse and natural human motion sequences based on textual descriptions constitutes a fundamental and challenging research area within the domains of computer vision, graphics, and robotics. Despite significant advancements in this field, current methodologies often face challenges regarding zero-shot generalization capabilities, largely attributable to the limited size of training datasets. Moreover, the lack of a comprehensive evaluation framework impedes the advancement of this task by failing to identify directions for improvement. In this work, we aim to push text-to-motion into a new era, that is, to achieve the generalization ability of zero-shot. To this end, firstly, we develop an efficient annotation pipeline and introduce *MotionMillion*—the largest human motion dataset to date, featuring over 2,000 hours and 2 million high-quality mo-

tion sequences. Additionally, we propose *MotionMillion-Eval*, the most comprehensive benchmark for evaluating zero-shot motion generation. Leveraging a scalable architecture, we scale our model to 7B parameters and validate its performance on *MotionMillion-Eval*. Our results demonstrate strong generalization to out-of-domain and complex compositional motions, marking a significant step toward zero-shot human motion generation. The code is available at <https://github.com/VankouF/MotionMillion-Codes>.

1. Introduction

Text-to-motion generation, which synthesizes diverse and temporally coherent human motions from natural language descriptions, poses a significant challenge in computer vision, graphics, and robotics. Despite notable advancements [1, 8–10, 14, 41, 63, 92, 94] in large-scale genera-

[†] Corresponding author

tive models for text, images, 3D, and videos—showcasing exceptional zero-shot capabilities—the motion domain [27, 72, 87, 99, 105] remains considerably behind. This discrepancy is attributed not to a lack of algorithmic innovation, but to inherent limitations in data scale and model architecture that hinder robust generalization, thereby constraining real-world applicability.

To address the generalization challenges faced by state-of-the-art methods such as MDM [72], MotionGPT [36], and MoMask [27], which are limited by dataset constraints in HumanML3D [25] and MotionX [45], earlier approaches [30, 71] aligned motion sequences with the image embedding space of CLIP [61] using rendered frames. While these methods somewhat alleviate data scarcity, they suffer from inherent modality mismatches, as static image embeddings fail to capture temporal dynamics, resulting in incoherent motions and limited compositional reasoning. Recent efforts [43, 52, 78, 89] have aimed to enhance generalization by scaling up motion generation models with larger datasets (e.g., 250 hours of motion in ScaMo [52] and one billion parameters in OMG [43]). These methods exhibit improved motion diversity and novel language alignment compared to previous methods [27, 36, 72], which were trained on more restricted datasets [25, 45]. Nonetheless, impeded by constraints related to model capacity and the inherent limitations of the datasets (e.g. quality, diversity, and size), the full potential of this scaling-up formulation remains largely unexploited, particularly when these text inputs entail long-term motion compositions and complex descriptions.

Therefore, we contend that achieving human-level motion generation necessitates a paradigm shift akin to the “*scaling hypothesis*”: sufficiently large and diverse high quality training data, combined with scaled model architectures, can unlock emergent zero-shot capabilities, especially performing the complex compositional motions. To this end, in this paper we explore three key components for zero-shot motion generation: 1) large-scale and high-quality motion dataset, 2) scalable and model architecture, 3) effective evaluation benchmark.

Firstly, we introduce MotionMillion, a large-scale motion dataset with comprehensive text annotations. We propose a novel and efficient motion annotation mechanism, including motion descriptions and high quality motion capturing, sourced from web-scale human motion videos. Our framework autonomously harvests human motion from unlabeled videos through kinematic regression, generates semantically rich captions using advanced vision-language models (e.g., gpt-4o [1]), and implements a multi-stage filtering process to eliminate scene cuts and static pose jitters. This meticulous curation results in a dataset comprising over **2000 hours** of high-quality text-motion pairs, encompassing over **2 million** motion sequences—**20 times** larger

than existing resources. By unifying annotations across extant datasets (e.g., HumanML3D [25], MotionX [45]), we establish a temporally coherent and compositionally diverse foundation for scaling.

Leveraging this dataset, we further explore the effective scalable model to accommodate our large-scale dataset. LLAMA, as a transformer decoder-only model, has significantly demonstrated its scalability in text generation tasks. In the motion field, it has been proven in Scamo [52] that a scaling law curve can be drawn with the increase in data volume. Therefore, 1) we first use Finite Scalar Quantization (FSQ) as **Efficient Motion Tokenization**: to discretely encode the motion data. This is a more stable and efficient way compared to VQ-VAE. However, we found that although the discretization method of FSQ can effectively encode motion data when the data scale is limited, due to the extremely large scale of our data, directly using the FSQ model will cause jitter in the reconstructed motion. We believe that this is because the information loss caused by the discretization method of FSQ becomes more serious as the data scale increases, leading the model to wrongly model high-frequency information. To solve this problem, we propose to use wavelet transformation to preprocess the motion data before inputting it into the reconstruction network. After obtaining the decoder outputs, we further utilize the inverse wavelet transformation to obtain the final reconstructed motions. By this means, we can effectively encode the motion data and reduce the jitter phenomenon caused by discrete information loss. After completing the discrete compression encoding, we further utilize the LLAMA architecture to implement 2) **Scalable Motion Generation**: We design a bidirectional transformer that jointly models text-motion cross-attention and autoregressive motion token prediction, enabling compositional motion synthesis. Starting from a 1B parameter base, we progressively scale model depth to the final 7B scale, observing emergent zero-shot capabilities. As shown in Fig. 1, our model could deal with various texts, especially can follow complex long texts.

To systematically assess the zero-shot generalization capabilities of models, we further introduce MotionMillion-Eval, a new benchmark comprising 126 diverse prompts across 7 categories, ranging from daily life scenarios to inhuman motions. Our evaluation focuses on three key aspects: text-motion alignment, motion smoothness, and physical feasibility of motions. Our findings indicate that the 7 billion parameter model, MotionMillion, successfully trained on the MotionMillion dataset, exhibits robust zero-shot generalization abilities. This advancement paves the way for advancing the motion generation task towards zero-shot applications.

Our contributions can be summarized as follows:

- We propose a high-quality annotation pipeline of human motions from video data and build MotionMillion,

a large-scale human motion dataset, which is currently the largest human motion with the highest quality, and its scale and diversity drive the research in the field of human motion towards zero-shot application.

- We propose to leverage the wavelet transformation to decrease the jitter phenomenon from FSQ, and we further scale our model to 7B parameters via an effective scalable architecture, demonstrating strong generalization for out-of-domain complex compositional motions.
- We built the MotionMillion-Eval benchmark according to industry standards, which is the first proposed evaluation that can be used for zero-shot capability verification.

2. Related Work

Text-aligned Human Motion Generation [2–4, 6, 7, 11–13, 15–18, 22, 23, 26–28, 30, 31, 33–36, 38, 46, 48, 51, 56, 57, 59, 64, 71, 72, 75, 76, 80, 81, 83–86, 88, 91, 93, 96, 98–102, 104, 106, 107] has progressed rapidly in recent years, benefiting substantially from advances in generative models [29, 67, 68, 74] and the expansion of large-scale datasets [25, 45, 90]. Although physics-based motion generation methods [21, 32, 54, 55, 77, 79, 97] can generate actions that are more in line with physical laws, text-align kinematic motion generation can possess higher flexibility. Methodologically, the introduction of GPT-like approaches [27, 36, 51, 99] and diffusion-based methods [13, 19, 72, 101, 102, 106] has substantially driven innovation in human motion generation. Meanwhile, KIT [58] and HumanML3D [25] have emerged as key benchmarks supporting text-driven motion generation. Nevertheless, the models tends to overfit the above datasets and lose the generalization capabilities.

Large Motion Model. Recent studies [43, 45, 90] seek to enhance generation quality by scaling dataset sizes. Concurrently, various works [36, 78, 82, 103, 104, 108] focus on enlarging model capacities, such as fine-tuning pre-trained large language models [36, 78, 82, 104, 108], although the resulting performance has often been suboptimal. LMM [103] implements a large diffusion model but suffers from slow training speed. ScaMo [52] try to scale the dataset size, motion vocabulary size, and the autoregressive model size and first to explore the scaling law in text-driven motion generation. Despite these advancements, current models exhibit limited generalization, thereby constraining their applicability in downstream tasks. We attribute this challenge primarily to the insufficient scale of current datasets. Therefore, in this work, we further expand both dataset sizes and model capacities, aiming to advance text-driven motion generation toward zero-shot scenarios.

3. MotionMillion Construction

3.1. Overview

Our dataset construction integrates human motion reconstruction from large-scale in-the-wild video sources and the re-aggregation of existing motion datasets. To enhance diversity and coverage, we incorporate multiple established datasets, including MotionX [45], InterHuman [44], Inter-X [90], BABEL [60], Fitness [45], PhantomDance [40], GDance [39], FineDance [42], HI4D [95], TRUMANS [37], and HumanSC3D [24]. To further scale data collection, we propose an efficient pipeline for reconstructing human motion from web-scale video sources.

Our methodology primarily focuses on full-body motion while omitting hand and facial expressions. To represent human motion, we extract SMPL[50] parameters, a widely adopted parametric model for human body articulation. As illustrated in Fig.2, our motion reconstruction framework consists of six key stages: which are **I**) Shot Segmentation, **II**) Human Detection, **III**) Bounding Box Confidence Filtering, **IV**) Transition Filtering, **V**) SMPL Motion Estimation, and **VI**) Motion Filtering.

3.2. Human Motion Reconstruction

In this section, we will illustrate the motion reconstruction process from web-scale human motion videos in detail.

Stage I Shot Segmentation. Raw video data often exhibits varying quality and frequent scene transitions, leading to artifacts in SMPL parameter estimation and adversely impacting downstream tasks. To address these challenges, we employ PySceneDetect to segment videos into single-scene clips, enforcing temporal coherence by restricting each clip to a maximum of 200 frames. Additionally, we use the Laplacian operator from OpenCV to evaluate image sharpness, selecting the sharpest frame as the initial frame of each clip. This preprocessing pipeline enhances the robustness of motion reconstruction while mitigating noise in the extracted SMPL parameters.

Stage II Robust Human Detection and Tracking. Accurate human detection and tracking are crucial for high-quality SMPL motion estimation. However, videos sourced from online platforms present two major challenges: (1) varying numbers of individuals present in each frame and (2) severe occlusions. To address these issues, we propose a **coarse-to-fine** approach that enhances the robustness of human detection and tracking, ensuring reliable motion estimation across diverse and complex scenarios.

During the coarse stage, we leverage Grounding DINO [49] and SAM2 [62] to fix the problem of variable human counts. Grounding DINO trains on an extremely large-scale dataset, and can detect any object based on the input text. SAM2 is a foundation model for solving promotable visual segmentation in images and videos. It can

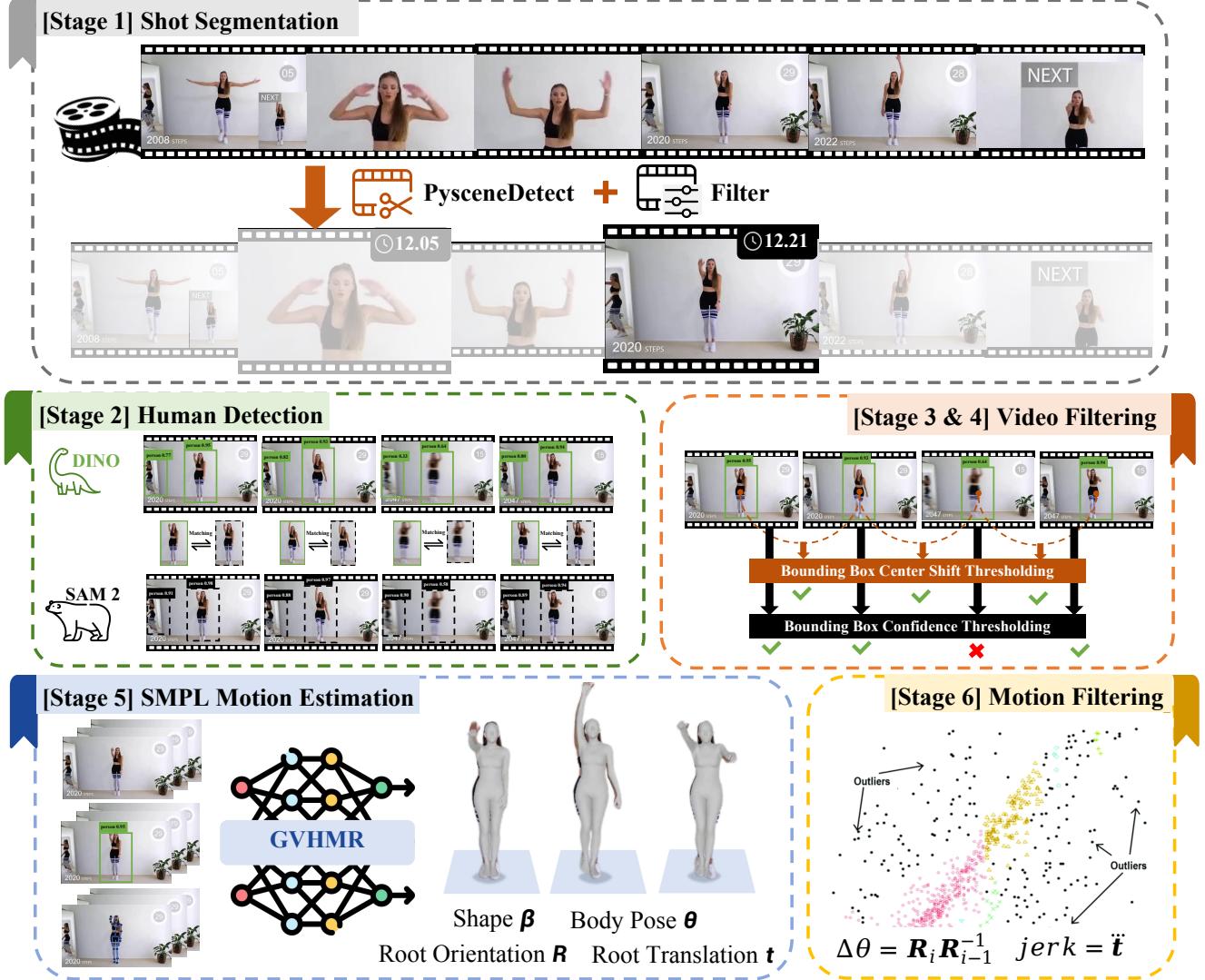


Figure 2. Data Construction Pipeline of MotionMillion. We can obtain high-quality human motion from a monocular video via our six processing stages, i.e. Shot Segmentation, Human Detection, Video Filtering, SMPL Motion Estimation and Motion Filtering.

use the bounding box of a person in the input image as a prompt to track that person’s identity information in subsequent frames. Specifically, **1)** Use Grounding DINO for high-confidence human detection. If the first frame lacks a person, scan subsequent frames until a valid detection. We empirically set the threshold at 0.85 to make sure the detected human body contains high quality. **2)** Feed the detected bounding box into SAM2, which propagates the mask across frames via prompt-based segmentation.

During the fine stage, we aims to solve the problem of severe occlusions. SAM2 demonstrates strong robustness in tracking individuals, even under occlusions, successfully preserving identity information. However, occlusions significantly degrade the accuracy of human keypoint detection, thereby impacting motion reconstruction quality. To

mitigate this, we leverage Grounding DINO’s confidence score as an indicator of detection reliability and introduce a refinement mechanism to filter low-quality bounding boxes: **1)** Calculating the IoU between SAM2 tracked boxes and Grounding DINO’s per-frame detections. **2)** Selecting candidates in Grounding DINO’s detection with IoU greater than 0.85, then choosing the box with minimal area deviation from the box tracked by SAM2. **3)** If the confidence score of the selected bounding box is greater than a certain threshold, it is considered a successful match; otherwise, it is considered a failed match. By matching the corresponding bounding boxes and filtering according to the confidence level, we can effectively alleviate the situation where a large number of occluded humans are detected due to the robustness of SAM2.



Figure 3. Overview of MotionMillion. This dataset exhibits extensive semantic and pose diversity, encompassing a broad spectrum of indoor and outdoor human motions.

Stage III & IV Bounding Box Confidence Filtering and Transition Filtering. We finish the bounding box confidence filtering during the fine stage of stage 2. While PySceneDetect is employed for preliminary shot segmentation, it struggles with scenarios where the background remains unchanged, but the subject undergoes sudden positional shifts. To address this limitation, we introduce an additional detection step for sudden position changes. We calculate the distance between the centers of the bounding boxes of detected humans in two consecutive frames. If it is greater than a certain threshold, it is considered that there is a sudden position change, and thus the video clip is divided into two parts.

Stage V Human Motion Reconstruction. To obtain high-quality human parameters, we used the GVHMR [65], which, as the state-of-the-art method in human motion reconstruction, could recover a much more realistic motion in both camera and world space. It first propose estimating human poses per-frame in a novel Gravity-View (GV) coordinate system. The GV system is defined by the world gravity and the camera view direction, which could naturally align with gravity and largely reduce the ambiguity in

defining the world coordinate system. It takes the bounding box, 2D key points, image features, and relative camera rotations as input, and leverage a transformer network to predict the SMPL parameters, including the root translation t , body pose θ , root orientation R , and shape β .

Stage VI Motion Filtering. PySceneDetect algorithm is limited in detecting sudden orientation changes in the foreground, particularly in scenarios where the background remains unchanged. While GVHMR demonstrates strong performance in human motion estimation, it remains susceptible to errors caused by camera motion-induced jitter, leading to inconsistencies in human body estimation. Therefore, to mitigate these issues, we integrate global orientation (R) and joint's position (J) estimated by GVHMR to effectively detect sudden orientation changes while filtering out jitter-related artifacts. To filter sudden orientation changes, we compute the transformation angle $\Delta\theta$ between two consecutive frames using the global orientation R , formulated as:

$$\Delta\theta = \text{Transform}(R_i R_{i-1}^{-1}) \quad (1)$$

, where i indicates the index of the verified frame and the Transform represents the function of transforming the ro-

Dataset	Clip Number	Hour	Text Diversity	Static Motion	Person	Scene		
						Indoor	Outdoor	RGB
KIT-ML	3911	11.2	1-3	no	single	yes	no	no
BABEL	13220	43.5	1	no	single	yes	no	no
HumanML3D	14616	28.6	1-3	no	single	yes	no	no
Motion-X	81084	144.2	1	no	single	yes	no	no
MotionBase	1M	-	-	yes	multi	yes	yes	yes
MotionMillion (ours)	2M	>2000	>20	no	multi	yes	yes	yes

Table 1. Statistics comparison between our MotionMillion and other motion datasets.

tation from matrix form into the axis-angle form. For jitter filtering, we introduce the *jerk* metric, which is sensitive to kinetic irregularities and can effectively indicate the motion smoothness [5]. The jerk is defined as the time derivative of acceleration, formulated as:

$$jerk = \ddot{J}_i \quad (2)$$

, where J represents the global position of different body joints. After obtaining the $\Delta\theta$ and *jerk* metrics for all of the video clips, we take the Isolation Forest [47] algorithm to identify the outliers for both metrics respectively, which can detect the moments when sudden orientation changes and jitters occur in the video clip in an unsupervised manner.

Our construction mechanism can generate precise and smooth motions. Finally, we standardize each motion to 30fps according to the frame rate of the corresponding raw video, obtaining our final motion.

3.3. Motion Caption.

Different from previous motion datasets (e.g. MotionX), our motion caption contains two steps: Motion Description and Description Augmentation.

Motion Description. We split the video according to the results obtained from the previous human motion reconstruction to obtain the corresponding video clip. Then, we input this clip to GPT-4o to obtain the description of the human action in the corresponding bounding box. If the video title contains a description of the corresponding semantics, it is input to GPT-4o together. At the same time, different from MotionX, in addition to the description of body parts, we also focus on prompting the GPT-4o model to describe the age, body characteristics, movement styles, emotions, and environments of the subject, which is critical during motion generation. All the prompts used are shown in the appendix.

Description Augmentation MotionX only describes the motion once, resulting in insufficient diversity of text descriptions of the same motion and subsequently significantly inhibits the model generalization ability. Therefore, after obtaining the motion description, we further prompt the model of LLAMA 3.1-8B [73] to rewrite the motion description 20 times without changing the original semantic meaning, so as to realize the description augmentation.

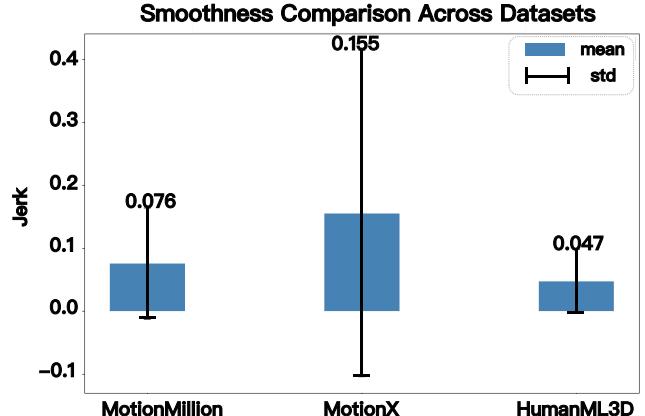


Figure 4. Jerk comparison across MotionMillion, MotionX, and HumanML3D. Our MotionMillion exhibits the lowest jerk values, indicating that it produces smoother motion.

3.4. Data Analysis

As presented in Table 1, our collected MotionMillion dataset comprises over 2,000 hours of human motion clips, encompassing more than 2 million motion sequences. Each sequence is ensured to exceed one second in duration, recorded at a frame rate of 30 frames per second. To enhance both the appearance and motion diversity, we curate a large collection of monocular videos from online sources, capturing a wide range of real-world scenarios. As depicted in Figure 3, our dataset encompasses various real-life settings, including indoor activities (e.g., performances, boxing) and outdoor movements (e.g., exercise, martial arts).

We further assess data quality from multiple perspectives, including pose diversity, semantic diversity (illustrated at the top of Figure 3), and motion smoothness (depicted in Figure 4). To evaluate motion smoothness, we compute the average jerk of the dataset, as formulated in Equation 2. The results indicate that the motion smoothness achieved through our data construction pipeline is significantly enhanced compared to MotionX and near to the overall smoothness of HumanML3D, thereby ensuring the high quality of our dataset.

Additionally, we visualize the top-2 principal components of body part poses and text features using t-SNE dimensionality reduction. The comparative analysis of pose distributions demonstrates that the diversity of poses within MotionMillion is on par with that of other datasets. However, in terms of semantic diversity, our dataset exhibits a significantly richer distribution compared to existing benchmarks. This is an expected outcome, as human motion often involves recurring pose patterns, yet distinct pose combinations yield semantically varied movements. Consequently, the increased semantic richness in our dataset presents challenges for smaller models to fully capture the underlying

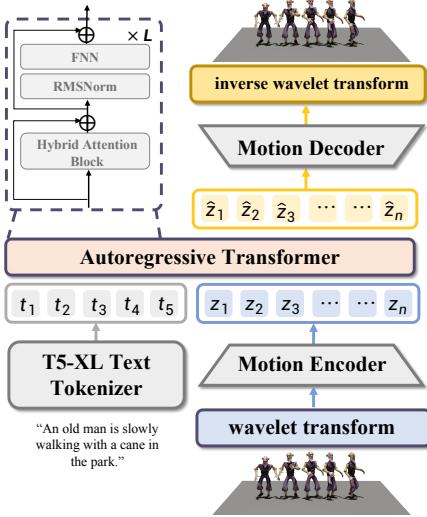


Figure 5. Overview of our scalable model architecture, which utilize FSQ as a motion tokenizer and an autoregressive transformer to generate the motion from the given text.

distribution, thereby stimulating further research in text-to-motion generation. In particular, our dataset underscores the necessity of scaling up model sizes and advancing the field towards zero-shot human motion generation.

4. Architecture

Building upon the constructed large-scale annotated motion dataset, we aim to train a foundational motion generation model capable of zero-shot motion generalization, especially the ability to generate complex compositional motions. Inspired by the successful scaling strategies observed in natural language and computer vision[1, 8, 9, 63], we adopt a discrete autoregressive architecture, as illustrated in Fig. 5. The proposed model consists of two key stages: Efficient Motion Tokenization and Scalable Motion Generation. 1) Efficient Motion Tokenization: This stage employs a finite scalar quantizer (FSQ) to learn discrete representations of human motion sequences, enabling efficient encoding of continuous motion data into a compact and structured format. 2) Scalable Motion Generation: Leveraging the LLAMA architecture, the model takes text inputs as prompts and scales from 1B to 7B parameters, facilitating high-capacity motion generation with strong generalization capabilities. This design enables the model to generate realistic and contextually coherent human motion sequences while maintaining scalability.

4.1. Motion Representations

We begin by introducing our motion representation. To mitigate errors introduced by the inverse kinematics process in the HumanML3D format while preserving redundant infor-

mation (e.g., velocity), we reformulate and refine the motion representation x^i in a manner consistent with previous work on character control [66, 69, 70]. Specifically, the i -th pose x^i is defined as a tuple comprising: root linear velocities ($\dot{r}^x, \dot{r}^z \in \mathbb{R}$) on the XZ-plane, root angular velocity $\dot{r}^a \in \mathbb{R}^6$ represented in 6D rotations, local joint positions $p^i \in \mathbb{R}^{3N}$, local velocities $v^i \in \mathbb{R}^{3N}$, and local rotations $r^i \in \mathbb{R}^{6N}$ relative to the root space, where N denotes the number of joints. Formally, this is expressed as:

$$x^i = \{\dot{r}^x, \dot{r}^z, \dot{r}^a, p^i, v^i, r^i\}.$$

A significant advantage of our representation is that it eliminates the need for an inverse kinematics process to obtain SMPL or BVH representations, as required in previous approaches. Moreover, our representation can be losslessly converted to relative rotations akin to those in SMPL. Additionally, because both the rotation and position components are derived from the same skeletal structure, they provide mutual regularization. Notably, the rotation component in the HumanML3D format is erroneous [20], which heavily hinders the applications of downstream tasks. Rather than discarding this flawed rotation component as done in [53], we undertake engineering corrections to rectify it. We hope our representation could correct previous mistakes and guide future development.

4.2. Efficient Motion Tokenization

We leverage Finite Scalar Quantization (FSQ), a codebook-free quantization mechanism that replaces distance-based codebook matching with deterministic discretization, enabling scalable and stable representation learning [52].

The vanilla FSQ architecture operates in three key steps. First, the latent vector z , produced by the motion encoder, is normalized via a sigmoid function to constrain its values within a bounded range $[0, 1]$. Next, each dimension of the normalized latent is discretized into L uniformly spaced integers using a rounding operation:

$$\hat{z} = Q(z) = \text{round}(f(z) \cdot (L - 1)), \quad (3)$$

, where L defines the number of quantization levels per dimension. This produces a discrete code $\hat{z} \in \{0, 1, \dots, L - 1\}^d$, with d denoting the latent dimension. Unlike VQ-VAE’s explicit codebook, FSQ implicitly defines a structured grid of L^d unique codes, eliminating the need to store physical embeddings while ensuring full code utilization.

The model is optimized solely via reconstruction loss:

$$\mathcal{L} = \|\mathbf{m} - \text{Dec}(\mathbf{z}_q)\|_2^2, \quad (4)$$

removing auxiliary losses (e.g., codebook commitment terms) required in VQ-VAE. This structured quantization paradigm provides a robust alternative to traditional VQ for representing complex motion data at scale.

Method	Dataset		
	HumanML3D	MotionX	MotionMillion
ScaMo [52]	63.3	84.1	88.9
Ours	41.9	57.4	45.5

Table 2. MPJPE of reconstruction comparison across different datasets, where ScaMo’s FSQ model and ours are trained on MotionUnion and MotionMillion, respectively.

	MPJPE↓	Mean Acc↓	Max Acc↓
GT	-	2.0	9.0
w/o wavelet	46.8	6.0	15.0
w/ wavalet	45.5	4.0	12.0

Table 3. Ablation on whether to use wavelet transformation during training the FSQ model, where Acc represents the acceleration.

Unlike the vanilla FSQ method, as shown in Fig. 5, we first use a wavelet transform before inputting the motion into the motion encoder, and after the motion decoder, we use the inverse wavelet transform to restore the motion. This can largely suppress the jitter problem caused by the information loss of discrete encoding.

4.3. Scalable Motion Generation

To effectively scale our model, we adopt a transformer-decoder-based architecture. The framework first compresses and encodes motion using a finite scalar quantizer (FSQ), while textual information is processed using the T5-XL large language model, which performs word-level encoding. The encoded motions and texts are denoted as $\{m_i\}_{i=1}^n$ and $\{T_i\}_{i=1}^w$, where n and w represent the number of encoded motion tokens and the word tokens.

Unlike standard causal attention mechanisms used in transformers, our approach employs a mixed attention strategy, where attention among words is bidirectional, whereas attention between motion sequences remains causal. As illustrated in Fig. 5, the encoded text and motion representations are fed into a series of stacked Hybrid Attention Blocks (HABs), with the final output directed to a classification head. Each HAB consists of: 1) Two RMS-Norm layers, which significantly mitigates the training instability. 2) One mixed attention module, which capture intricate relationships between text and motion features, and 3) one feedforward network(FFN) module, which could fuse the feature from a channel aspect. To optimize the model, we apply cross-entropy loss to the logits produced by the final classification head, formulated as:

$$\mathcal{L} = - \sum_{i=1}^n \log p(\hat{m}_i | m_{<i}, T_1, \dots, T_w) \quad (5)$$

This approach ensures robust motion-text alignment while maintaining scalability and training stability.

Method	FID↓	R@1↑	R@2↑	R@3↑
ScaMo	89.0	0.67	0.81	0.87
Ours-1B	31.3	0.74	0.87	0.92
Ours-3B	10.8	0.79	0.91	0.94
Ours-7B	10.3	0.79	0.90	0.94

Table 4. Quantitative comparison of ScaMo and our models of different sizes on MotionMillion.

5. Experiments

5.1. FSQ Reconstruction Comparison

Reconstruction Comparison with Different Datasets.

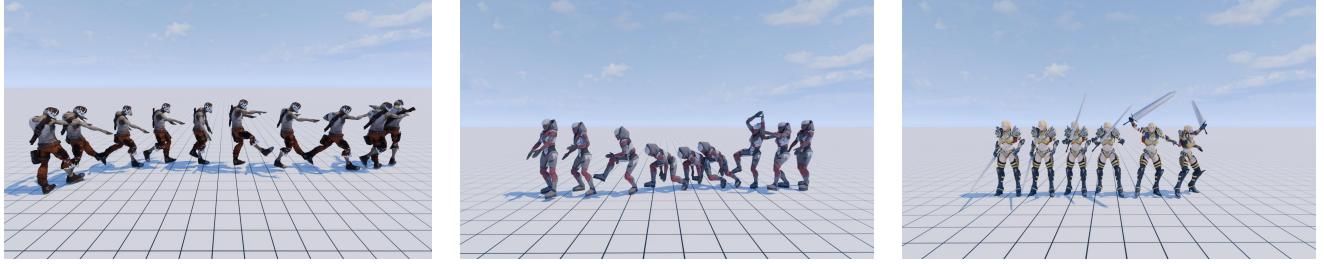
We first compare the reconstruction performance of different datasets after conducting FSQ training with the same number of parameters on both our dataset and the MotionUnion dataset. As shown in Tab. 2, our FSQ model is trained on the training set of MotionMillion, while the FSQ model from Scamo is trained on the MotionUnion dataset. We observe that the FSQ model trained on the MotionMillion training set, using the same number of parameters, achieves the best performance across HumanML3D, MotionX, and MotionMillion.

Ablation Study Wavelet Transformation. As shown in the Tab. 3, we exhibit the reconstruction results via MPJPE metrics and jitter by calculating the acceleration of the reconstructed motion.

We find that training vanilla FSQ (a.k.a. w/o wavelet) leads to a large deviation between the acceleration of the motion reconstructed by the model and that of the ground truth (GT), which means there is a significant data jitter. We analyze that this may be because the discretized compression inevitably causes information loss, resulting in a deviation in the modeling of high-frequency information, thus leading to severe jitter. Therefore, we use the wavelet transform to transform the input motion, and after using the wavelet transform, the deviation between the acceleration of the motion reconstructed by the model and that of the GT is significantly reduced. At the same time, the MPJPE of the model also gains a little better improvement.

5.2. Quantitative Comparison on Different Models

We followed the previous method [51], and trained an evaluator on our dataset to assess the impact of different model scales on the generation ability. As shown in Tab. 4, we can find that our model significantly surpasses the ScaMo performance. Besides, with the increase of the model size, FID and R-precision are gradually improved. However, we found that when the model size improved from 3B to 7B, it was difficult to effectively reflect the difference between the two models from the metrics. This is also in line with the current dilemma of generative models, that is, metrics often can not fully reflect the capabilities of the model.



The zombie shuffles forward with slow, unsteady steps, its feet dragging heavily against the ground. Its decayed arms stretch outward, fingers curled stiffly as if reaching for something unseen.

A lean-built individual executing a fluid sequence of movements that begins with a pivot, a spinning sword stroke, a transfer of weight to their back leg, a lunge with a strong thrust, and a smooth retreat into a stable stance, demonstrating their mastery of the movement.

With a katana in hand, the person freezes, then swiftly rotates their body while raising the sword before delivering a swift slash, their fluid motion and intense focus evident throughout the movement.

Figure 6. Our model demonstrates robust performance in generating coherent motions from complex compositional textual descriptions.

Model	Text Alignment	Physical Plausibility	Motion Smoothness
MDM [72]	195.5	478.5	416.5
MotionGPT [36]	170	497	501.5
T2M-GPT [99]	207	495.5	500
ScaMo-3B [52]	226.6	477.5	494
Ours-1B	170.3	497	501
Ours-3B	238.6	496	499.5
Ours-7B	261	495.5	501

Table 5. Average human evaluation results under the aspect of Text Alignment, Physical Plausibility, and Motion Smoothness on MotionMillion-Eval between Different Models.

Ours-7B vs. ScaMo-3B (win/tie/lose)	Annotator 1	Annotator 2	Annotator 3
Overall (126)	45/49/32	32/76/18	47/53/26
Art/Dance (2)	1/0/1	0/2/0	1/0/1
Combat (6)	3/2/1	2/2/2	3/2/1
Communication (20)	5/13/2	3/16/1	8/11/1
Daily life (74)	24/27/23	21/40/13	25/31/18
Non-human behavior (2)	2/0/0	0/2/0	0/1/1
Sports (16)	6/6/4	5/9/2	7/6/3
Work (6)	4/1/1	1/5/0	3/2/1

Table 6. Detailed comparison between ours-7B and ScaMo-3B. The green cells indicate that our model outperforms competing approaches, the yellow cells represent a tie, and the white cells denote that our model underperforms.

Therefore, in the following sections, we further propose the MotionMillion-Eval benchmark to evaluate zero-shot generation capabilities of different models.

5.3. Zero-shot Potential Verification

Benchmark and Metric. We introduce MotionMillion-Eval, a novel benchmark designed to assess the quality of text-to-motion generation models through human verification. This benchmark comprises 126 prompts derived from real-world industrial standards, covering various motion categories, including daily life, work, arts, communication, combat, sports, dance, and nonhuman behavior. We evaluate different text-to-motion models with three human evaluation dimensions on MotionMillion-Eval: 1) Text Alignment, 2) Motion Smoothness, and 3) Physical Plausibility. The scoring details for the three dimensions are provided in the supplementary materials.

Comparisons Between Different Models. As shown in Tab. 5, We evaluate multiple models on MotionMillion-Eval. Compared with the ScaMo-3B model, our 3B model significantly outperforms ScaMo in terms of Text Alignment, Physical Plausibility, and Motion Smoothness. This indicates that compared to the MotionUnion dataset used for training ScaMo, the model trained with our dataset has stronger generalization performance. In addition, as the model scale increases (ranging from 1B to 7B), the effect in the dimension of Text Alignment is significantly improved. This shows that the scale and diversity of our data can effectively support the expansion of the model scale, thus better paving the way for realizing zero-shot applications.

Besides, we find that for the two metrics of physical feasibility and motion smoothness, expanding the model size does not bring a significant increase. We analyze that the reason for this is that these two metrics do not need to measure the alignment degree between motion and text, but only need to focus on the quality of motion itself. Therefore, we believe that these two metrics are highly correlated with the motion quality of the dataset itself. Further, when we compare the MDM, T2M-GPT, and ScaMo-3B models, we find that our model can be comparable to the former two in terms of Physical Plausibility and Motion Smoothness, and is significantly better than the ScaMo-3B model. Therefore, this further shows that the quality of MotionMillion we constructed can significantly reach almost the same level as HumanML3D and is significantly better than MotionX.

Detailed Comparisons Between Ours-7B and ScaMo-3B Models. We compared the generative capabilities of both models on diverse instruction categories, with three professional annotators voting on the better outputs for identical text prompts. As shown in Table 6, our model matches ScaMo-3B in the Art/Dance and Non-human behavior subsets but surpasses it in all other areas, demonstrating the enhanced diversity of motions achievable by our dataset.

5.4. Qualitative Results

We further present visualization results, as shown in Fig. 6. Our model demonstrates: 1) the ability to comprehend ab-

stract concepts, effectively recognizing and generating the characteristic walking posture of a zombie; and 2) strong instruction-following capabilities, accurately interpreting long text descriptions and producing corresponding combined motions. These results indicate that MotionMillion, the largest-scale dataset we propose, has the potential to advance text-to-motion generation into the zero-shot era.

6. Conclusion

In this paper, we take the first step toward advancing human motion generation into the zero-shot era. We begin by analyzing the limited generalization ability of existing methods, attributing it to the constrained size of current datasets. To address this, we propose an efficient data annotation mechanism, establishing the largest annotated human motion dataset. Furthermore, to assess the zero-shot capabilities, we introduce MotionMillion-Eval, a dedicated evaluation benchmark. Building on this foundation, we successfully scale our model to 7B parameters using a scalable architecture, achieving state-of-the-art performance on the benchmark and demonstrating strong zero-shot capabilities. We believe this work lays a crucial foundation for advancing zero-shot applications in motion generation.

References

- [1] OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgium, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory De-careaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ibai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrej Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Jun-tang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. 2023.
- [2] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, pages 5915–5920, 2018.
- [3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728, 2019.
- [4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and GÜl Varol. Teach: Temporal action composition for 3d hu-

- mans. In *3DV*, pages 414–423, 2022.
- [5] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–469, 2024.
- [6] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *CVPR*, pages 457–469, 2024.
- [7] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *VR*, pages 1–10, 2021.
- [8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.
- [9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.
- [10] Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, et al. Dancetogether! identity-preserving multi-person interactive video generation. *arXiv preprint arXiv:2505.18078*, 2025.
- [11] 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.
- [12] Ling-Hao Chen, Shunlin Lu, Wenxun Dai, Zhiyang Dou, Xuan Ju, Jingbo Wang, Taku Komura, and Lei Zhang. Pay attention and move better: Harnessing attention for interactive motion generation and training-free editing. *arXiv preprint arXiv:2410.18977*, 2024.
- [13] 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 *CVPR*, pages 18000–18010, 2023.
- [14] Weihao Cheng and Ying Shan. Learning layout generation for virtual worlds. *Computational Visual Media*, 10(3):577–592, 2024.
- [15] Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: Language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024.
- [16] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, pages 9760–9770, 2023.
- [17] Minyue Dai, Jingbo Wang, Ke Fan, Bin Ji, Haoyu Zhao, Junting Dong, and Bo Dai. Towards synthesized and editable motion in-betweening through part-wise phase representation. *arXiv preprint arXiv:2503.08180*, 2025.
- [18] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pages 390–408. Springer, 2024.
- [19] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *ECCV*, 2024.
- [20] GitHub discussion. rotation discussion. <https://github.com/EricGuo5513/HumanML3D/issues/26>, 2023-03-04.
- [21] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C-ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH Asia 2023*, pages 1–11. 2023.
- [22] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024.
- [23] Yuming Feng*, Zhiyang Dou*#, Ling-Hao Chen, Yuan Liu, Tianyu Li, Jingbo Wang, Zeyu Cao, Wenping Wang, Taku Komura, and Lingjie Liu. Motionwavelet: Human motion prediction via wavelet manifold learning. *arXiv preprint arXiv:2411.16964*, 2024.
- [24] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1343–1351, 2021.
- [25] 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.
- [26] 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, 2022.
- [27] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910, 2024.
- [28] Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. Amd: Autoregressive motion diffusion. In *AAAI*, pages 2022–2030, 2024.
- [29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, pages 6840–6851, 2020.

- [30] 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 SIGGRAPH*, 2022.
- [31] Yiheng Huang, Hui Yang, Chuachen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. *ACM MM*, 2024.
- [32] Yiming Huang, Zhiyang Dou, and Lingjie Liu. Modskill: Physical character skill modularization. *ICCV 2025*, 2025.
- [33] Bin Ji, Ye Pan, Yichao Yan, Ruizhao Chen, and Xiaokang Yang. Stylevr: Stylizing character animations with normalizing flows. *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- [34] Bin Ji, Ye Pan, Zhimeng Liu, Shuai Tan, Xiaogang Jin, and Xiaokang Yang. Pomp: Physics-constrainable motion generative model through phase manifolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22690–22701, 2025.
- [35] Bin Ji, Ye Pan, Zhimeng Liu, Shuai Tan, and Xiaokang Yang. Sport: From zero-shot prompts to real-time motion generation. *IEEE Transactions on Visualization and Computer Graphics*, 2025.
- [36] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2024.
- [37] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024.
- [38] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *CVPR*, pages 2151–2162, 2023.
- [39] Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8673–8682, 2023.
- [40] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *AAAI*, pages 1272–1279, 2022.
- [41] Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhi-hong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*, 2024.
- [42] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *ICCV*, pages 10234–10243, 2023.
- [43] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–493, 2024.
- [44] Han Liang, Wenqian Zhang, Wenzhan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *IJCV*, pages 1–21, 2024.
- [45] 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. *NeurIPS*, 2024.
- [46] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018.
- [47] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.
- [48] Jinpeng Liu, Wenzun Dai, Chunyu Wang, Yiji Cheng, Yansong Tang, and Xin Tong. Plan, posture and go: Towards open-world text-to-motion generation. *ECCV*, 2024.
- [49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [50] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.
- [51] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *ICML*, 2024.
- [52] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenzun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. *arXiv preprint arXiv:2412.14559*, 2024.
- [53] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*, 2024.
- [54] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *2024 International Conference on 3D Vision (3DV)*, pages 1498–1507. IEEE, 2024.
- [55] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhs: Unified synthesis of physical human-scene interactions through task tokenization. *CVPR 2025*, 2025.
- [56] Mathis Petrovich, Michael J Black, and Gü̈l Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497, 2022.
- [57] 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 *CVPRW*, pages 1911–1921, 2024.

- [58] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.
- [59] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *RAS*, 109:13–26, 2018.
- [60] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021.
- [61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [62] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [64] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024.
- [65] 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 2024 Conference Papers*, pages 1–11, 2024.
- [66] Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with autoregressive motion diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–14, 2024.
- [67] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015.
- [68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.
- [69] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6):178, 2019.
- [70] Sebastian Starke, Ian Mason, and Taku Komura. Deep-phase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (ToG)*, 41(4):1–13, 2022.
- [71] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, pages 358–374, 2022.
- [72] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2022.
- [73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.
- [75] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *ECCV*, 2024.
- [76] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. In *ECCV 2024*, pages 37–54. Springer Nature Switzerland, 2024.
- [77] Wenjia Wang, Liang Pan, Zhiyang Dou, Zhouyingcheng Liao, Yuke Lou, Lei Yang, Jingbo Wang, and Taku Komura. Sims: Simulating human-scene interactions with real world script planning. *ICCV 2025*, 2025.
- [78] 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.
- [79] Yinhui Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17540–17549, 2025.
- [80] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *NeurIPS*, pages 14959–14971, 2022.
- [81] Zan Wang, Yixin Chen, Baoxiong Jia, Puahao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *CVPR*, pages 433–444, 2024.
- [82] 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:2405.17013*, 2024.
- [83] 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. *arXiv preprint arXiv:2503.15451*, 2025.
- [84] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024.

- [85] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *ICLR*, 2024.
- [86] Zhenyu Xie, Yang Wu, Xuehao Gao, Zhongqian Sun, Wei Yang, and Xiaodan Liang. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In *AAAI*, pages 6252–6260, 2024.
- [87] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *ICCV*, pages 2228–2238, 2023.
- [88] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *ICCV*, pages 2228–2238, 2023.
- [89] Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations. *arXiv preprint arXiv:2410.13790*, 2024.
- [90] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, pages 22260–22271, 2024.
- [91] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *CVPR*, pages 1759–1769, 2024.
- [92] Qun-Ce Xu, Tai-Jiang Mu, and Yong-Liang Yang. A survey of deep learning-based 3d shape generation. *Computational Visual Media*, 9(3):407–442, 2023.
- [93] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *arXiv preprint arXiv:2405.13659*, 2024.
- [94] Yuhang Yang, Fengqi Liu, Yixing Lu, Qin Zhao, Pingyu Wu, Wei Zhai, Ran Yi, Yang Cao, Lizhuang Ma, Zheng-Jun Zha, et al. Sigman: Scaling 3d human gaussian generation with millions of assets. *arXiv preprint arXiv:2504.06982*, 2025.
- [95] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023.
- [96] Chengjun Yu, Wei Zhai, Yuhang Yang, Yang Cao, and Zheng-Jun Zha. Hero: Human reaction generation from videos. *arXiv preprint arXiv:2503.08270*, 2025.
- [97] Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. *arXiv preprint arXiv:2505.02094*, 2025.
- [98] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, pages 16010–16021, 2023.
- [99] 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 *CVPR*, pages 14730–14740, 2023.
- [100] Jiaxu Zhang, Xin Chen, Gang Yu, and Zhigang Tu. Generative motion stylization of cross-structure characters within canonical motion space. In *ACM MM*, 2024.
- [101] 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*, 2023.
- [102] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 2024.
- [103] 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. *arXiv preprint arXiv:2404.01284*, 2024.
- [104] 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 *AAAI*, pages 7368–7376, 2024.
- [105] Qiu Zhou, Manyi Li, Qiong Zeng, Andreas Aristidou, Xiaojing Zhang, Lin Chen, and Changhe Tu. Let’s all dance: Enhancing amateur dance motions. *Computational Visual Media*, 9(3):531–550, 2023.
- [106] 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. *ECCV*, 2024.
- [107] 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 and high-quality motion generation. In *ECCV 2024*, pages 18–38. Springer Nature Switzerland, 2024.
- [108] 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.

Go to Zero: Towards Zero-shot Motion Generation with Million-scale Data

Supplementary Material

7. Overview

In the following, we provide additional implementation details, including the data distribution and prompts in Sec. 8, finegrained scoring criteria in MotionMillion-Eval in Sec. 9, and 126 prompts used by MotionMillion-Eval in Sec. 10. Various generations of out-domain and complex compositional long motions are shown in our demo video.

8. Data Distribution and Prompts

We further demonstrate the data distribution of motion length, motion velocity, and motion diversity in Fig. 7. We also provide the prompt used during captioning the motions in the web-scale human videos and text rewrite in the inference stage in Fig. 8.

9. Scoring Criteria Details of MotionMillion-Eval

The scoring criteria details for each dimension (Text Alignment, Motion Smoothness, and Physical Plausibility) are defined as follows:

Text Alignment (TA). Score = 4: The generated motion is fully aligned with the textual prompt, accurately depicting all specified elements and details. Score = 3: The motion generally corresponds to the prompt, though minor discrepancies may be present in certain details. Score = 2: The motion exhibits clear misalignment with the prompt, with significant omissions or deviations from the described content. Score = 1: The generated motion is entirely inconsistent with the prompt, displaying substantial inaccuracies in key scenes or actions.

Motion Smoothness (MS). Score = 4: The motion is highly fluid and natural, with smooth and seamless transitions between movements. Score = 3: The motion is generally smooth, though minor unnatural artifacts may occasionally appear in specific segments. Score = 2: The motion lacks fluidity, exhibiting noticeable discontinuities or stuttering. Score = 1: The motion appears highly unnatural, with frequent stuttering and abrupt transitions that disrupt coherence and comprehensibility.

Physical Plausibility (PP). Score = 4: The generated motion adheres to real-world physical laws, accurately simulating object interactions, lighting, shadows, and collision effects. Score = 3: Multiple instances of physically implausible motion, lighting inconsistencies, or unrealistic interactions are observed, though the primary actions maintain a degree of coherence. Score = 2: The generated motion exhibits substantial violations of physical laws, with unre-

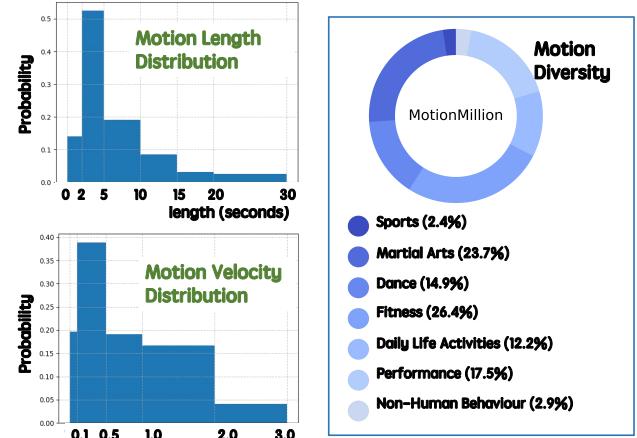


Figure 7. Data Distributions of MotionMillion.

Caption Prompt

[Task Overview]: You will be shown different frames uniformly sampled from a human motion video, along with a general description of the video. Based on these, identify the actions performed by the person within the bounding box and provide a detailed description of their movements.

[Required Information]: 1. Body Part Involvement: Describe the main limbs involved in the movement. 2. Action Sequence: Provide a clear, coherent sentence structure that captures the sequence of movements, including joints, body parts, and their logical relationships. 3. Temporal Progression: Focus on how different body parts move over time in relation to one another. Use appropriate transitions to show the movement sequence.

[Optional Information]: 1. Physical Characteristics: Age, body shape, and gait of the person, if relevant. 2. Movement Range or Style: Describe the range or style of movement, e.g., “walking like an elderly person” or “swift, like a martial artist.” 3. Scene Description: Location context, e.g., “indoor,” “outdoor,” etc. 4. Emotional or Attitudinal Cues: Any emotions or attitudes expressed, such as “appears focused and determined.”

The descriptions should be concise, accurate, and based on the provided video frames and the overall scene description.

Augmentation Prompt

You are a helpful assistant. You can rewrite the following sentence and do not change the meaning of the sentence. You can change the subject and the way of description, but do not alter the original meaning.

[Requirements:] 1. Do not change the meaning of the sentence. 2. Does not change the logical relationship of the sentences. 3. The rewritten sentence should be more concise and clear, and write in a pair of brackets. Example: Given the sentence: A man is walking forward for a few steps. The rewritten sentence is: {[A man is walking stightly.]}

Now rewrite the following sentence for 20 times: {sentence}

Figure 8. Prompt used during captioning the motions in the web-scale human videos and text rewrite in inference stage. The prompt includes sections for Caption Prompt (Task Overview, Required Information, Optional Information), Augmentation Prompt (Instructions, Requirements), and a note about not changing the meaning or logical relationships while making the sentence more concise and clear.

10. Prompts in MotionMillion-Eval

- . We show all the prompts in the figures below.

An obese middle-aged male security guard, walking and looking around	Clap hands	Clap once
The boy clutched the flowers tightly with both hands, hiding them behind his back, his body taut and eyes fixed ahead. As the girl approached, his face instantly lit up with a smile, and he quickly strode forward, swiftly extending his right hand to offer the flowers.	Drink water	Turn left
The man stood in the downpour. Upon seeing the woman, his head jerked to the left with a pained look of avoidance, and his body leaned back slightly.	Open the door	Bow lightly
The woman approached the man. Her right hand trembled as she raised it, intending to touch his face.	Snap fingers	Kick forward
The reserved man and the shy woman sat side by side on the edge of the bed. After a furtive glance at the woman, the man coughed lightly, his body gradually edging closer. His hands rubbed on his knees, his movements somewhat awkward.	Bow	Twirl hair
The woman's hands were intertwined, nervously twisting the corner of her clothing, her gaze cast downward.	A teenager jogs slowly in the park, occasionally checking the phone	Knock twice
In the cluttered kitchen, the slightly plump man frantically flipped the pan with his right hand while his left hand scrambled to grab the seasonings. After knocking over the salt shaker, oil splattered in the pan. His eyes widened, his mouth gaped open, and he jumped on the spot with both feet. The spatula flew out of his hand, and he stumbled, looking utterly disheveled.	A young woman quickly folds clothes in a messy living room	Wave goodbye
The boss was speaking in a measured and serious manner, with his left hand behind his back and his right hand occasionally gesturing. Suddenly, upon hearing a sound of flatulence, his expression changed abruptly, and his right hand froze in mid-air.	Pick up a pen from the floor and place it on the table	Spin around
A man of average build who looked lost was walking along the street when a giant pie suddenly hit his head. He clutched his head with both hands and squatted down.	A furious swordsman grips his blade tightly, stomps forward with an angry roar, then slashes diagonally at an invisible foe	Toss a ball
The woman gazed out at the vast ocean, her body swaying as she slowly and heavily made her way towards the precipice.	A confident performer in a flashy costume strikes a dramatic pose, then leaps into a high-flying cartwheel across the stage	Jump high
A strong man in deep sorrow rushed to the scene, running while reaching out and shouting with a voice filled with despair, his steps faltering.	A young boy trudging through knee-deep snow, occasionally pausing to catch his breath	Snap fingers

The emaciated woman sat on the floor, her hands wrapped around her knees as she trembled, her head buried in her arms, sobbing, her body curled up.	A woman practicing yoga, gracefully transitioning from a downward dog position to a cobra pose	Drink water
The woman ran in small quick steps, gradually slowing down, standing on her tiptoes to wave, gazing into the distance with tearful eyes.	students practicing a comedic skit, overacting their gestures and laughing at each other's mistakes	Tie shoelaces
The tall and capable woman adjusted her collar in front of the mirror, took a deep breath, straightened her posture, and smiled confidently. With her right hand holding a bag and her left hand opening the door, she walked out with a light step, her posture upright.	A thin man frantically searching for his keys in a sandstorm, shielding his face with one arm and reaching around with the other	Punch forward
The athletes warmed up, their feet alternating in quick jumps, their hands clenched into fists swinging back and forth. After the whistle blew, they shot off like arrows, their feet rapidly alternating, their arms swinging with power, sprinting towards the finish line with all their might.	A middle-aged couple greeting each other warmly after a long day, wrapping their arms around each other in a tight hug	Slide left
The tall detective held a flashlight, moving slowly with a slight lean forward, turning his head to observe his surroundings. Upon spotting blood, he quickly crouched down, his left hand supporting his knee, and with his right hand, he directed the flashlight's beam, his gaze focused and thoughtful.	A zombie slowly dragging its feet forward, arms outstretched, letting out a low groan	Shake fists
The robust expedition leader held a map in his hand, occasionally looking up to observe the surrounding environment, moving slowly and cautiously. His left hand slightly raised to signal the team to halt, his right index finger pressed to his lips in a gesture for silence. His body was taut, and he strained to listen to any movements around them.	A robot spinning its torso 360 degrees, scanning the environment with glowing eyes, then extending a mechanical arm to pick up an object	Grab handle
A female college student was walking while reading a book when she was suddenly blinded by a dazzling light. She used her right hand to shield her eyes, and then was knocked to the ground by an unidentified object.	Shuffle sideways	Push door
A young girl is skipping rope in the playground.	Gently pat a dog's head	Stomp foot
An old man is slowly walking with a cane in the park.	A grandpa showing a child how to fish by a lake, casting the line then patiently waiting	Twist torso
A strong athlete is lifting heavy weights in the gym.	A teacher writing on the blackboard, pausing occasionally to ask questions	Wink quickly

A cute toddler is crawling on the floor.	Someone standing in the desert, shading their eyes from the sun and scanning the horizon	Bend knees
A middle-aged woman is practicing yoga on a mat.	A chef briskly chopping vegetables, occasionally wiping sweat from his brow	Shrug shoulders
A professional dancer is performing a ballet solo.	A woman nervously tapping her foot while waiting in line, looking at her watch repeatedly	Rest head
A basketball player is dribbling and shooting.	Flip hair back confidently, resting hand on hip	Tap foot
A construction worker is hammering nails.	A furious boxer hitting a punching bag repeatedly, sweat flying with each powerful strike	Look up
A schoolboy is running for the school bus.	A content farmer hoeing the field, wiping his forehead with the back of his hand	Crouch down
A female gymnast is doing somersaults on the balance beam.	A young woman wearing headphones, bobbing her head to the rhythm and tapping her fingers on the table	Pull rope
A waiter is carrying a tray of dishes.	A friend casually leaning against a wall, crossing one leg over the other and scrolling on their phone	Bow deeply
A skateboarder is doing tricks in the skate park.	A young couple having a heated argument in the living room, arms flailing as voices rise	Cross arms
A firefighter is climbing a ladder to rescue people.	Cautiously slide open a window, peering outside with curiosity	Duck quickly
A surfer is riding a big wave.	A man meticulously polishing his car, wiping down every inch with a cloth and stepping back to admire the shine	Rub neck
A tailor is sewing clothes with a sewing machine.	Throw a paper airplane across the room with a flick of your wrist	Scratch chin
A hunter is stalking prey in the forest.	A tall athlete performing a slam dunk on a basketball hoop, shouting in triumph	Walk slowly
A hairdresser is cutting a customer's hair.	Casually flick dust off your shoulder	Point up
A cyclist is racing in a mountain bike competition.	A florist arranging a bouquet, gently snipping stems and adjusting petals	Kick side
A pianist is playing a passionate piece on the piano.	A street performer juggling three brightly colored balls, smiling as the crowd gathers	Arch back

Jump rope	A shy child timidly stepping forward to receive an award, hands clasped together and head lowered	Greet politely
Stand still	A martial artist practicing a high roundhouse kick, exhaling sharply	Cross legs
Raise both hands	A scientist adjusting a microscope carefully, squinting into the eyepiece	Shake head
Kick a ball	A teenage boy throwing his backpack onto a couch and stretching out with a tired groan	Gather items
Wave hello	A woman sipping tea while flipping through a magazine, occasionally glancing out the window	Open door