MotionLab: Unified Human Motion Generation and Editing via the Motion-Condition-Motion Paradigm

Ziyan Guo[1] Zeyu Hu[2] De Wen Soh[1] Na Zhao[1†]

[1] Singapore University of Technology and Design, Singapore [2] LIGHTSPEED, Singapore

ziyan_guo@mymail.sutd.edu.sg, {dewen_soh, na_zhao}@sutd.edu.sg, hzy9724@gmail.com

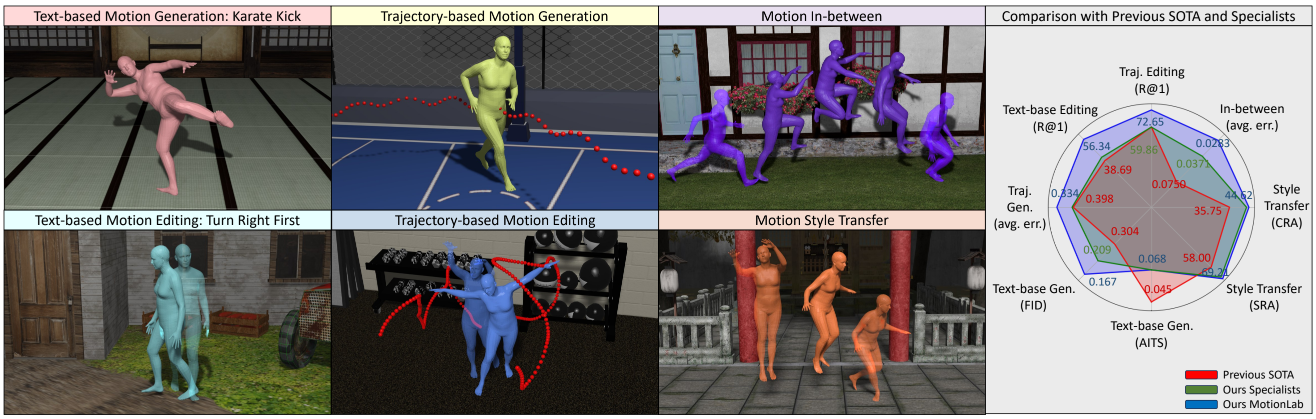


Figure 1. Demonstration of our MotionLab's versatility, performance and efficiency. Ours specialists refer to the proposed framework tailored for specified tasks. Previous SOTA refer to multiple models, including MotionLCM [12], OmniControl [63], MotionFix [6], CondMDI [11] and MCM-LDM [55]. All motions are represented using SMPL [39], where transparent motion indicates the source motion or condition, and the other represents the target motion. **More qualitative results are available in the website and appendix.**

Abstract

*Human motion generation and editing are key components of computer vision. However, current approaches in this field tend to offer isolated solutions tailored to specific tasks, which can be inefficient and impractical for real-world applications. While some efforts have aimed to unify motion-related tasks, these methods simply use different modalities as conditions to guide motion generation. Consequently, they lack editing capabilities, fine-grained control, and fail to facilitate knowledge sharing across tasks. To address these limitations and provide a versatile, unified framework capable of handling both human motion generation and editing, we introduce a novel paradigm: **Motion-Condition-Motion**, which enables the unified formulation of diverse tasks with three concepts: source motion, condition, and target motion. Based on this paradigm, we propose a unified framework, **MotionLab**, which incorporates rectified flows to learn the mapping from source motion to target motion, guided by the specified conditions. In MotionLab, we introduce the 1) MotionFlow Transformer to enhance conditional generation and editing without task-specific modules; 2) Aligned Rotational Position Encoding to guarantee the time synchronization between source motion and target motion; 3) Task Specified Instruction Modulation; and 4) Motion Curriculum Learning for effective multi-task learning and knowledge sharing across tasks. Notably, our MotionLab demonstrates promising generalization capabilities and inference efficiency across multiple benchmarks for human motion. Our code and additional video results are available at: https://diouo.github.io/motionlab.github.io/.*

1. Introduction

Human motion is a crucial component of computer vision, with applications spanning game development, film production, and virtual reality [21, 59]. With the advancements of generative diffusion models [13, 28, 53], human motion generation has garnered considerable attention, aiming at generating human motion aligned with the input conditions, such as text [58, 59, 66] and trajectory (*i.e.*, joints' coordinates) [4, 5, 12, 18, 50, 54, 57, 63, 67]. Concurrently, to maximize the utility of motion assets within industry settings, significant efforts have been dedicated to motion editing tasks, including motion style transfer [1, 24, 29, 55, 69].

†Corresponding author: Na Zhao.

Method	text-based generation	text-based editing	trajectory-based generation	trajectory-based editing	in-between	style transfer
MDM [59]	✓	×	×	×	–	×
MLD [9]	✓	×	×	×	×	×
OmniControl [63]	✓	×	✓	×	–	×
MotionFix [6]	–	✓	×	×	–	×
CondMDI [11]	✓	×	✓	×	✓	×
MCM-LDM [55]	×	×	×	×	–	✓
MotionGPT [30]	✓	–	×	×	✓	×
MotionCLR [8]	✓	–	×	×	–	–
Ours	✓	✓	✓	✓	✓	✓

Table 1. Summary of methods focusing on motion generation and editing. ✓ indicates that the method has been trained for the task, × indicates that the method fails to implement, and – indicates that the method has not been trained but can implement in a zero-shot manner.

As summarized in Table 1, current research in this domain mainly develops task-specific solutions, forcing practitioners to train multiple models for human motion generation and editing, which is inefficient and impractical. Although several studies [16, 41, 52, 64, 68, 70, 71] have attempted to unify motion-related tasks, they merely consider different modalities as generation conditions, leading to limited editing capabilities and insufficient fine-grained trajectory control. Moreover, these approaches overlook the intrinsic links between motion generation and editing, thereby hindering potential knowledge sharing. In contrast, a well-designed unified framework can exploit the large volumes of multi-task data to surpass specialist models through effective cross-task representation learning. Motivated by this prospect and inspired by the success of large language models in unifying NLP tasks [2, 14], we pose the following question: ***Can human motion generation and editing be effectively unified within a single framework?***

In response to this question, it is essential to design an elegant and scalable paradigm. Hence, we propose a novel paradigm: **Motion-Condition-Motion**. This paradigm is built upon three concepts – *source motion*, *condition*, and *target motion*. Concretely, the target motion is predicted by the source motion and specified conditions in this Motion-Condition-Motion paradigm. For any human motion generation task, the source motion can be treated as none, and the target motion must align with the provided conditions. For any human motion editing task, the target motion is derived from the source motion based on the conditions. By unifying these tasks within this elegant and scalable paradigm, this framework can be seamlessly extended to various human motion tasks and scaled across diverse datasets. Given that human motions are inherently tied to their semantics, trajectories, and styles in practical applications, we aim to unify several key tasks under this framework. These tasks include *text-based motion generation and editing* [6, 19, 58, 59, 66], *trajectory-based motion generation and editing* [12, 25, 63], *motion in-between* [11, 26] and *motion style transfer* [55, 69], as illustrated in Figure 1.

Despite the proposed paradigm, several significant challenges remain in balancing versatility, performance, and efficiency: 1) Unifying various tasks inevitably introduces additional modalities, while each modality may involve multiple tasks. A naive solution, like adopting multiple cross-attention mechanisms for each task in generation-unified frameworks [16, 68], is suboptimal. 2) More sampling time is required for certain tasks (*e.g.*, trajectory-based motion generation and motion in-between [63, 69]), as existing methods in these areas involve task-specific posterior guidance [10] during inference to improve conditional guidance. 3) Time asynchrony between the source motion and target motion may arise due to the limited scale of the paired editing dataset and the use of implicit positional encoding [6, 9, 11, 63]. 4) Most importantly, naively integrating various motion generation and editing tasks into a single framework could lead to task conflicts and catastrophic forgetting, impairing the framework's overall performance.

To address these challenges, we propose a novel generative framework, termed **MotionLab**, built upon rectified flows [37, 38] and MM-DiT [15], as illustrated in Figure 2. Rectified flows are particularly well-suited for MotionLab since they are designed to implement the optimal transport between source and target distributions, naturally aligning with the Motion-Condition-Motion paradigm. Furthermore, human motion must adhere to skeletal kinematics. Consequently, the distribution of valid target motions is highly restricted, and individual target motions may correspond to multiple conditions (*i.e.*, many-to-one), which suggests that a shared transport map can be efficiently transferred across tasks. In contrast to MM-DiT, our proposed Motion Flow Transformer (MFT) encompasses more modalities, including source motion, target motion, text, trajectory, and style. Within the MFT, each modality is assigned a dedicated path, and comprehensive interaction among modalities is facilitated via joint attention. This architecture enables MFT to advance conditional generation and editing capabilities without necessitating task-specific modules or posterior guidance for particular tasks. To ensure temporal synchronization between source and target motions, we incorporate an Aligned Rotational Position Encoding into MFT, which explicitly aligns tokens at corresponding frames between the source and target sequences. Moreover, to enable adaptation of a single modality to various tasks, we introduce Task Instruction Modulation, which flexibly embeds various tasks into the MFT. To seamlessly integrate diverse tasks, we propose a curriculum-inspired training strategy, termed Motion Curriculum Learning, based on an easy-to-hard training principle.

In this paper, motion generation and editing are decomposed into combinations of modalities through the Motion-Condition-Motion paradigm. These modalities are subsequently represented via each modality's paths within the MFT, learning inter-modal interactions through the joint attention, while adapting individual modalities to different tasks through Task Instruction Modulation. By implementing a curriculum learning from single to multiple, from simple (*e.g*., source motion and trajectory) to complex (*e.g*., text and style) modalities, the spatial knowledge inherent in 3D representations can be effectively transferred to the latter modalities since the modalities of the former can represent the latter. Through these designs, we validate MotionLab on multiple benchmarks, demonstrating superior versatility, performance, and efficiency compared to baselines across various human motion generation and editing tasks.

2. Related Work

Motion Generation and Editing. Motion generation can be classified based on input conditions. Among these, text-based motion generation is one of the most compelling areas [9, 20–23, 30, 33, 35, 40, 44, 46, 58, 59, 61, 65, 66], as it trains models to comprehend the semantics of text and generate corresponding pose sequences. To address the fine-grained requirements of practical applications, trajectory-based motion generation has been proposed [12, 32, 51, 63, 67], where specific motion properties, such as joints reaching designated positions at specified times, are defined. Additionally, motion in-between [11, 30, 45, 48, 59] focuses on generating complete motion sequences given key poses at keyframes. To enable in-place editing of human motion [6, 19], MotionFix [6] introduces text-based motion editing using paired source and target motions. We extend this approach to trajectory-based motion editing by substituting text with joint trajectories. Meanwhile, style plays a crucial role in human motion, leading to motion style-transfer [1, 29, 55, 69]. However, the aforementioned methods concentrate solely on specific tasks, rendering them impractical for real-world applications. Moreover, they overlook the intrinsic connections across different human motion tasks and fail to facilitate knowledge sharing among these tasks. In contrast, our unified framework enhances performance on data-scarce editing tasks through multi-task learning.

Unified Frameworks for Human Motion. There are also some efforts in existing methods that try to unify tasks related to human motion. One line of work [6, 30, 31, 34, 36, 41, 61, 62, 72] focuses on motion understanding, such as motion captioning or describing human motion in images and videos. Yet, these approaches often rely on GPT-like structures, which require a large amount of training resources and GPU memory. In addition, they fail to provide fine-grained control (*e.g*., trajectory-based generation and editing) over motion, which is crucial in practical applications. Another line of effort [3, 16, 41, 52, 64, 68, 70, 71] highlights generating motion based on more modalities, such as music and speech. However, these approaches only integrate more modalities into one model and cannot flexibly edit motion, which can cause them to suffer from multi-task learning and limit their scope of use. The closest to our work are FLAME [33] and MotionCLR [8]. However, FLAME does not support style transfer and precise text-based editing like "move faster", and MotionCLR does not support trajectory-based generation and editing, requiring cumbersome manual adjustments to the attention.

3. Preliminary: Rectified Flows

Flow-based generative methods [15, 17, 37, 38, 42, 47] have recently received significant attention due to their generalizability and efficiency compared to diffusion models. Specifically, these methods directly regress the transport vector field between the source distribution p_1 and target distribution p_0 with the straightest possible trajectories and sample by the corresponding ordinary differential equation (ODE) [60]. Among these methods, rectified flows [37, 38] aim to learn a trajectory from source data x_0 to target data x_1, which can be formulated as $x_t = \varphi(x_0, x_1, t)$, and the velocity field v_t of the trajectory x_t can be defined by:

$$v_t = \frac{dx_t}{dt} = \frac{\partial \varphi_t(x_0, x_1, t)}{\partial t}, t \in [0, 1] \tag{1}$$

Once we have learned this velocity field v_t, we can get x_0 from any x_1 by numerically integrating:

$$x_{t-\frac{1}{N}} = x_t - \frac{1}{N} v_\theta(t, x_t) \tag{2}$$

where N is the discretization number of the interval [0,1]. Hence, rectified flows v_θ are trained to predict v_t by given x_t and t, and the training objective can be represented as:

$$\mathcal{L}_{RF}(\theta) = \int_0^1 \mathbb{E}_{(x_0,x_1)\sim(p_0,p_1)}[||v_\theta(t, x_t) - v_t||_2^2]dt \tag{3}$$

4. Motion-Condition-Motion

To unify the tasks of human motion generation and editing, we propose the paradigm of Motion-Condition-Motion. As shown in Table 2, all these tasks are unified by three concepts: *source motion*, *condition*, and *target motion*.

Motion Generation. For the motion generation tasks, including *text/trajectory-based* generation and *motion in-between*, the source motion can be treated as none, with

Task	Source Motion	Condition	Target Motion
unconditional generation	$\emptyset$	$\emptyset$	✓
masked reconstruction	masked source motion	$\emptyset$	source motion
reconstruction	complete source motion	$\emptyset$	source motion
text-based generation	$\emptyset$	text	✓
trajectory-based generation	$\emptyset$	text/joints' coordinates	✓
motion in-between	$\emptyset$	text/poses in keyframes	✓
text-based editing	✓	text	✓
trajectory-based editing	✓	text/joints' coordinates	✓
style transfer	✓	style motion	✓

Table 2. Structuring human motion tasks within our Motion-Condition-Motion paradigm.

the target motion aligning to the corresponding conditions. For instance, in *text-based generation*, the generated motion should align with the semantics of the provided text, such as "karate kick" illustrated in Figure 1. *Masked reconstruction*, as a specific motion generation task, requires the target motion to align with the masked source motion in the specified frames without relying on additional conditions. Notably, the *unconditional generation* (given zero frames) and *reconstruction* (given all frames) are special cases of masked reconstruction, thus these three tasks can share the same task instruction as described in Section 5.2.

Motion Editing. For motion editing, the source motion must be provided, and the target motion is derived from the source motion based on the specified conditions. In the case of *text-based motion editing*, the generated motion should originate from the source motion, with modifications applied only to the specified parts as dictated by the provided text, such as "use the opposite leg". For *trajectory-based editing*, the source motion should be aligned with the given joints' coordinates, ensuring that the specified joints in the source motion are accurately moved to the designated positions within the specified frames. In *motion style transfer*, the generated motion should adopt the style of the style motion while preserving the semantics of the source motion.

Remarks. In particular, *trajectory-based motion generation* and *motion in-between* are highly similar, as they both aim to ensure that specific joints reach designated positions at specific times. Their primary difference is that the former is sparse in space (*i.e.*, joints) but dense in time, whereas the latter is dense in space (*i.e.*, joints) but sparse in time. To efficiently share the parameters and learned representations between the two tasks, we unify their conditions into a single condition. Meanwhile, *masked reconstruction* is also similar to these two tasks. However, while these two tasks only include the coordinates of joints, the source motion also encompasses the velocity and angular velocity of joints. Therefore, they represent different modalities, and masked reconstruction constitutes a distinct task.

5. MotionLab

Based on our proposed Motion-Condition-Motion paradigm, we introduce a unified framework named **MotionLab**, as illustrated in Figure 2(a). The core of MotionLab is the **MotionFlow Transformer (MFT)** (Sec. 5.1), inspired by MM-DiT [15], which leverages rectified flow to map source motion $M_S \in \mathbb{R}^{N_S \times D}$ to target motion $M_T \in \mathbb{R}^{N_T \times D}$ based on the corresponding condition C for each task.

To enable task differentiation, we propose **Task Instruction Modulation** (Sec. 5.2), where a task-specific instruction $I \in \mathbb{R}^{1 \times 768}$ extracted from the CLIP [49] is also input into MFT alongside M_S, M_T, and C. At each timestep t, MFT is trained to predict velocity field v_t, which is derived via linear interpolation between target motion M_T and Gaussian noise $\epsilon \in \mathbb{R}^{N_T \times D}$.

For effective multi-task training, we adopt **Motion Curriculum Learning** (Sec. 5.3) which organizes tasks hierarchically to facilitate learning. Once trained, MotionLab can map M_S to M_T based on the specified C, by predicting v_t in descending order of timestep t as described in Sec. 3.

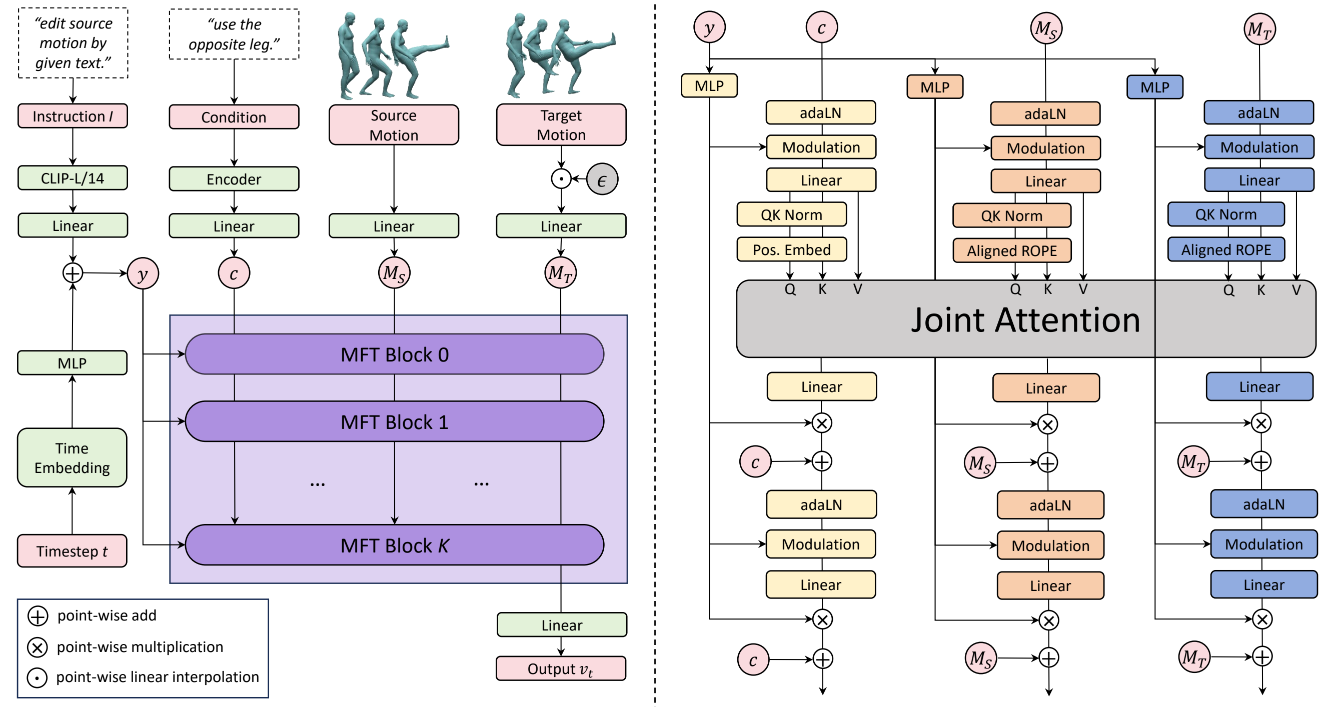
5.1. MotionFlow Transformer

As shown in the Figure 2 (b), MotionFlow Transformer contains three key components: *Joint Attention* to interact tokens from different modalities; *Modality Path* for distinguishing tokens from different modalities and extracting their representations, and *Aligned ROPE* for position encoding of modalities with time information.

Joint Attention. We first adopt the joint attention mechanism [15], through which tokens from different modalities can interplay with each other. Specifically, all these tokens will be projected to the query, key, and value representations, and then will be concatenated into a sequence of orderly tokens. Subsequently, these orderly tokens are applied by the attention operation, whose output is again split into corresponding tokens of different modalities.

Modality Path. While the joint attention can interact with tokens from different modalities, there is still a need to differentiate between different tokens. In addition to the QKV projection and FeedForward Network (FFN) in the attention mechanism, as used in MM-DiT, our MFT incorporates the adaptive Layer Normalization (adaLN) and a modulation mechanism [43] for each modality, enhancing conditional generation and editing capabilities.

Aligned Rotational Position Encoding. Considering that the use of absolute position encoding in existing methods [9] can weaken the temporal alignment between source motion and target motion due to the limited scale of paired datasets, we adopt a relative position encoding method, ROtational Position Encoding (ROPE) [56]. ROPE explicitly embeds the relative distances between tokens, preserving temporal relationships more effectively. Instead of naively applying a 3-dimensional ROPE to distinguish source motion, target motion, and conditions with time information (*e.g.*, trajectory), we propose Aligned ROPE, which encodes these components with appropriate temporal information using a 1-dimensional ROPE. This design avoids the



(a) Overall Architecture of MotionLab

(b) Detail of MotionFlow Transformer (MFT) Block

Figure 2. Illustration of our MotionLab and the detail of its MotionFlow Transformer (MFT).

confusion caused by 3-dimensional ROPE, where distances between tokens within a modality can interfere with cross-modality distances, ensuring better temporal alignment.

5.2. Task Instruction Modulation

MM-DiT implements a modulation mechanism that enhances text-to-image generation through the incorporation of textual embeddings (*e.g.*, "a photo of a dog") as modulation signals. However, within our unified framework, various tasks necessitate the integration of multiple modalities, and critically, identical modalities may require distinct representational forms across different tasks. This complexity renders approaches such as learned task tokens (*e.g.*, [TASK]) or one-hot encoding vectors inadequate for managing arbitrary numbers and combinations of modalities.

Recognizing the inherent flexibility of natural language, we leverage textual representations acquired by foundation models (*e.g.*, CLIP) to effectively differentiate identical modalities across disparate tasks. For instance, we utilize the textual embedding of "*edit source motion by given style*" to facilitate the adaptation of source motion to style transfer. This approach, while conceptually straightforward, provides remarkable effectiveness in enhancing system flexibility and scalability, thereby enabling seamless extension to diverse tasks involving multiple modalities.

5.3. Motion Curriculum Learning

To achieve effective multi-task learning and facilitate knowledge sharing between tasks, we propose an easy-to-hard hierarchical training strategy inspired by curriculum learning [7]. Specifically, new tasks are sequentially introduced into the training based on their difficulty, guided by the following assumptions: 1) The fewer modalities a task involves, the simpler the task; 2) Editing tasks are easier than generating tasks, as only the conditional difference between source motion and target motion needs to be learned; 3) The more specific the conditional information (*e.g.*, source motion) provided, the simpler the task becomes. The importance of these three criteria decreases in order. Guided by the easy-to-hard training principle, the training process in MotionLab is divided into two stages: *self-supervised pre-training* and *supervised fine-tuning*.

Pre-training. Intuitively, the reconstruction of masked source motion is the easiest task. Hence, we first train the model based on the masked source motion, independent of the conditions. This approach allows the model to learn prior motion representations independent of conditions, thereby generalizing to different tasks. Following MoMask [23], we randomly mask from zero frames to all frames. This flexible strategy provides tasks of varying difficulty levels, avoiding overfitting on simple tasks (all frames) and mode collapse on difficult tasks (zero frames).

Furthermore, this strategy seamlessly performs source motion reconstruction (*i.e.*, all frames) and unconditional training (*i.e.*, zero frames), which is crucial for Classifier-Free Guidance (CFG) [27]. Unlike MoMask, which masks all joints in a single frame simultaneously due to its discrete tokens, we extend masked pre-training to randomly mask joint trajectories to enhance the understanding of in-between and trajectory-based tasks. Specifically, we pre-train MotionLab using these three tasks (*i.e.*, masked source motion reconstruction, trajectory-based generation without text, and in-between without text) for 1,000 epochs.

Fine-tuning. In the supervised fine-tuning stage, we train MotionLab on tasks in an easy-to-hard sequence. Specifically, a new task is introduced into training every 200 epochs in the following order: ① text-based generation, ② style-based generation (an auxiliary task for training the modality path of the style, not our primary goal), ③ trajectory-based editing (without text), ④ text-based editing, ⑤ style transfer, ⑥ motion in-between and trajectory-based generation, ⑦ trajectory-based editing. This progressive learning strategy ensures effective adaptation and knowledge sharing across tasks. Particularly, ① and ② are the simplest tasks because they only include one modality, whereas others include at least two modalities. Among tasks involving two modalities, ③, ④, and ⑤ take priority over ⑥ since they are editing tasks. Additionally, as text is less specific than trajectory but more specific than style, the order is ③, ④, and ⑤.

To mitigate catastrophic forgetting, previous tasks are trained with new tasks, based on the probability derived from the FID of the last evaluation. However, the FID scales for different tasks vary due to their differing difficulty levels. Consequently, we use the percentage change compared to the previous evaluation as the probability, which encourages the model to re-learn forgotten tasks or tasks that it has not yet fully mastered. To support classifier-free guidance, we also train the model to unconditionally generate and reconstruct the complete source motion. Empirically, in this stage, a 5% probability is allocated for unconditional generation, 5% for reconstructing the complete source motion, 45% for previous tasks, and 45% for the new task.

In summary, this training strategy has many advantages: 1) it enables our framework to adapt to various tasks; 2) it seamlessly supports CFG during inference; 3) it allows flexible management of the training process to avoid retraining due to errors. Meanwhile, this training strategy, from single modality to multiple modalities, can be considered as first learning the representation of each modality separately and then learning the representation of the interaction between multiple modalities, which can be distinguished by the Task Instruction Modulation. Furthermore, by prioritizing the introduction of spatial conditions (*i.e.*, source motion and trajectory), this strategy can share the model's understanding between them and abstract conditions (*i.e.*, text and style), as the latter conditions can be represented by the former.

Method	FID↓	R@3↑	Diversity→	MM Dist↓	MModality↑	AITS↓
GT	0.002	0.797	9.503	2.974	2.799	-
T2M [21]	1.087	0.736	9.188	3.340	2.090	**0.040**
MDM [59]	0.544	0.611	**9.559**	5.566	<u>2.799</u>	26.04
MotionDiffuse [66]	1.954	0.739	11.10	2.958	0.730	15.51
MLD [9]	0.473	0.772	9.724	3.196	2.413	0.236
T2M-GPT[65]	**0.116**	0.775	9.761	3.118	1.856	1.124
MotionGPT [30]	0.232	0.778	<u>9.528</u>	3.096	2.008	1.240
CondMDI [11]	0.254	0.6450	9.749	-	-	57.25
MotionLCM [12]	0.304	0.698	9.607	3.012	2.259	<u>0.045</u>
MotionCLR [8]	0.269	**0.831**	9.607	**2.806**	1.985	0.830
Ours	<u>0.167</u>	<u>0.810</u>	9.593	<u>2.830</u>	**2.912**	0.068

Table 3. Evaluation of *text-based motion generation* on HumanML3D [21] dataset. The models in bold are the optimal models, and the models in underline are the sub-optimal models.

Method	Joints	FID↓	R@3↑	Diversity→	Foot skate ratio↓	Average Error↓	AITS↓
GT	-	0.002	0.797	9.503	0.000	-	-
GMD [32]	pelvis	0.576	0.665	9.206	0.101	<u>0.1439</u>	137.0
PriorMDM [51]	pelvis	0.475	0.583	9.156	-	0.4417	19.83
OmniControl [63]	pelvis	<u>0.212</u>	0.678	9.773	<u>0.057</u>	0.3226	39.78
MotionLCM [12]	pelvis	0.531	**0.752**	<u>9.253</u>	-	0.1897	**0.035**
Ours	pelvis	**0.095**	<u>0.740</u>	**9.502**	**0.007**	**0.0286**	<u>0.133</u>
OmniControl [63]	all	0.310	0.693	9.502	0.061	0.0404	76.71
Ours	all	0.126	0.765	9.554	0.002	0.0334	0.134

Table 4. Evaluation of *trajectory-based motion generation* on HumanML3D [21] dataset.

Method	Condition	generated-to-target retrieval				Average Error↓	AITS ↓
		R@1↑	R@2↑	R@3↑	AvgR ↓		
GT	-	100.0	100.0	100.0	1.00	-	-
TMED* [6]	text	38.69	50.61	62.23	4.15	-	26.57
Ours	text	56.34	70.40	77.24	3.54	-	0.16
TMED* [6]	trajectory	60.01	73.33	82.69	2.67	0.129	30.56
Ours	trajectory	72.65	82.71	87.89	2.20	0.027	0.19

Table 5. Evaluation of *text-based and trajectory-based motion editing* on MotionFix [6] dataset. TMED* mean that we re-implement the models since the original models are in the skeleton of SMPL format, while ours is in HumanML3D format.

6. Experiments

Datasets. To evaluate the text-based motion generation, the trajectory-based motion generation, motion in-between, and motion style transfer, we leverage the HumanML3D [21] dataset, which comprises 14,646 motions and 44,970 motion annotations. To evaluate the text-based and trajectory-based motion editing, we utilize MotionFix [6] dataset, which is the first dataset for text-based human motion editing, including 6,730 motion pairs.

Evaluation Metrics. We evaluate our framework using the following metrics: 1) To evaluate *text-based motion generation*, following the [9], we adopt the FID to evaluate the distribution gap between the generated and original motions; Diversity to calculate the corresponding variance between motions; R-precision (R@K) to measure the proximity of the generated motion to the text or motion; Foot skating

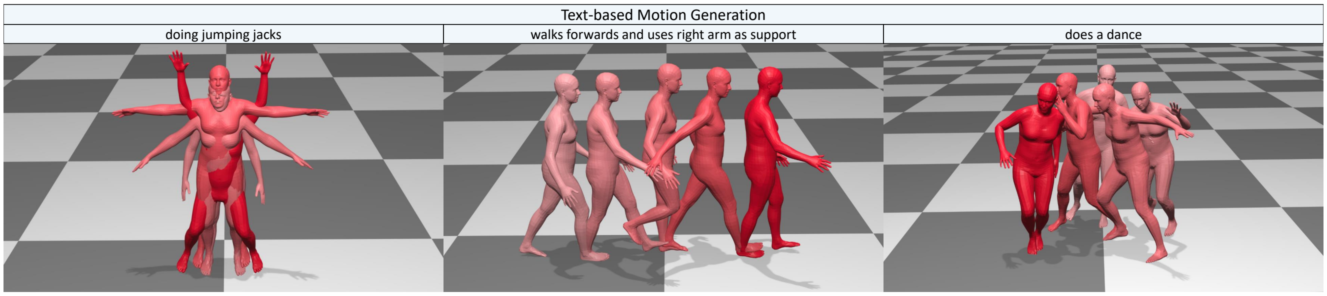


Figure 3. Qualitative results on the text-based motion generation. For clarity, as time progresses, motions transit from light to dark colors.

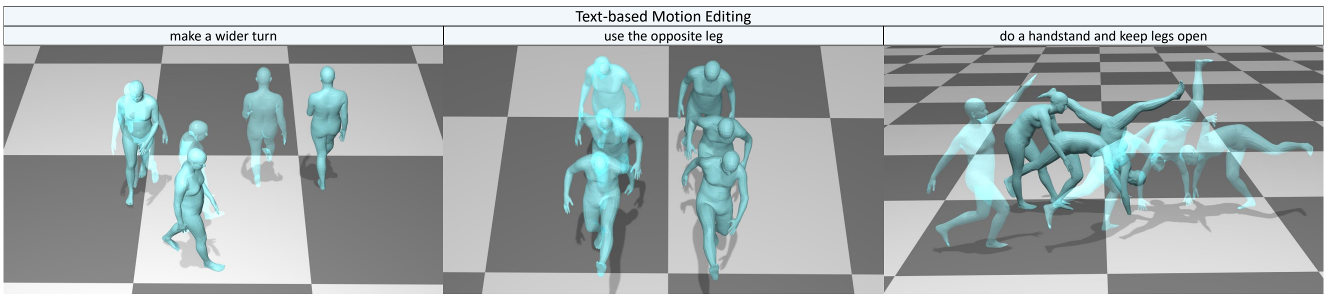


Figure 4. Qualitative results on text-based motion editing. The transparent motion is source motion, and the other is the generated motion.

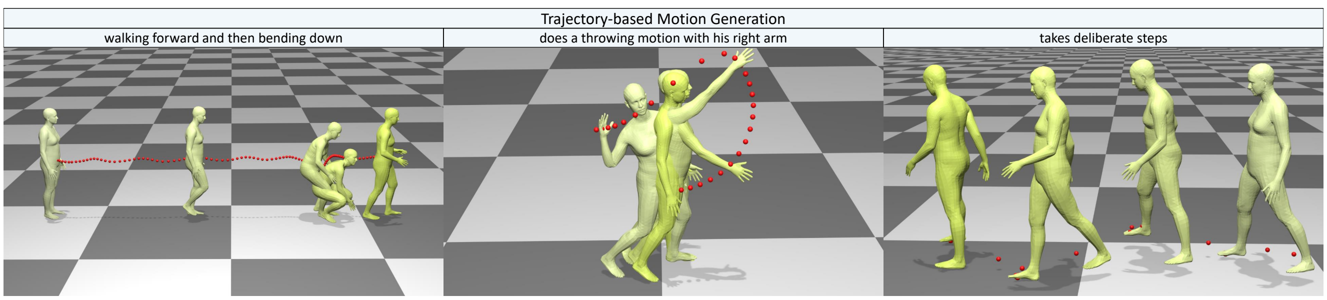


Figure 5. Qualitative results on the trajectory-based motion generation. The red balls are the trajectory of the pelvis, right hand and foot.

ratio to evaluate the physical plausibility of motion; Multimodal Distance (MM Dist) calculates the distance between motions and texts. We also introduce Average Inference Time per Sample (AITS) measured in seconds to evaluate the inference efficiency; 2) To evaluate *trajectory-based motion generation* and *motion in-between*, following [63], we adopt the Average Error to measures the mean distance between the generated motion locations and the keyframe locations; 3) To evaluate *text-based and trajectory-based motion editing*, following [6], we adopt the AvgR to measure the success rate of retrieval from edited motion to target motion; 4) To evaluate *motion style transfer*, following [55], we adopt the Style Recognition Accuracy (SRA) and Content Recognition Accuracy (CRA) to measure the stylistic and content accuracy of the generated motion; Trajectory Similarity Index (TSI) to evaluate the trajectory preservation from source motion.

Implementation Details. In order to fairly compare our model with other models, motions from all datasets have been retargeted into one skeleton following HumanML3D format with 20 fps, where the number of joints J is 22, and the dimension of motion feature D is 263. The learning rate is set to be 1×10^{-4}. The timesteps are set to 1,000 for training and 50 for inference. Our models are trained by four RTX 4090D with each batch of 64 for 4 days. To ensure a fair comparison, the AITS of all models are recalculated using one RTX 4090D.

6.1. Quantitative Results

Overall Performance. As shown in Table 3 to Table 5, MotionLab demonstrates promising performance across *all benchmarks*[*], underscoring the effectiveness of our framework's design. Notably, as MotionLab is a unified framework without task-specific designs, it must balance versatility, performance, and efficiency.

Specifically, as shown in Table 3 and Figure 6, MotionLab achieves superior performance (lowest FID, which is the key metric for generation tasks) with relatively fast

[*]Due to space limitations, we include the quantitative results on motion in-between and motion style transfer in the supplementary material.

Method	text gen. (FID)	traj. gen. (avg. err.)	text edit (R@1)	traj. edit (R@1)	in-between (avg. err.)	style transfer (CRA)	style transfer (SRA)
w/o rectified flows	0.301	0.0359	54.38	69.21	0.0289	42.20	63.96
w/o MotionFlow Transformer	0.483	0.0447	51.26	65.34	0.0349	35.36	53.83
w/o Aligned ROPE	0.253	0.0886	45.39	61.99	0.0756	42.23	56.59
w/o task instruction modulation	0.223	0.0401	55.96	70.01	0.0288	40.55	63.91
w/o motion curriculum learning	1.956	0.1983	28.56	36.61	0.1682	29.51	34.23
Ours specialist models	0.209	0.0398	41.44	59.86	0.0371	43.53	67.55
Ours	**0.167**	**0.0334**	**56.34**	**72.65**	**0.0283**	**44.62**	**69.21**

Table 6. **Ablation studies of key components of MotionLab on each task**. Refer to the text for the detailed configuration of each variant.

inference time (third-lowest AITS). For trajectory-based tasks (Table 4) and the motion in-between task, MotionLab achieves lower average error. We believe these improvements stem from the effectiveness of masked pre-training and Aligned ROPE, which ensures spatial and temporal synchronization between the trajectory and target motion.

6.2. Qualitative Results

As shown in the Figure 3 from Figure 5, our framework presents its powerful capabilities to generate motion aligned with the conditions and edit source motion based on the condition, demonstrating its versatility and performance. For more visualization results, please kindly refer to the supplementary and project website.

6.3. Ablation Studies

We perform several ablation experiments* on our framework to validate the designs in MotionLab and report the results in Table 6: the 1^{st} variant replaces rectified flows with diffusion models; the 2^{nd} variant uses a regular transformer (*i.e.*, without modulation mechanism and adopting cross-attention) instead of MFT. The 3^{rd} variant uses the implicit 1D-learnable encoding instead of Aligned ROPE; The 4^{th} variant does not adopt the Task Instruction Modulation; the 5^{th} variant directly learns all tasks based on their FID compared to the last evaluation. Additionally, we use the same model to train specialist models for each task, denoted as 'our specialist models' in Table 6.

As can be seen from the results, the removal of motion curriculum learning markedly diminishes model performance across all tasks, underscoring its pivotal role in facilitating knowledge transfer between tasks. Meanwhile, our unified framework outperforms our specialist models in all tasks, potentially due to the knowledge sharing of motion curriculum learning. These phenomena can also be attributed to the strategy's capacity to enable the model to integrate its comprehension of spatial conditions (*e.g.*, source motion, trajectory, and intermediate states) with abstract conditions (*e.g.*, text and style), given that the latter can be partially represented by the former. Furthermore, as shown in Table 6, Aligned ROPE is essential for space-related tasks, significantly reducing the average error. It effectively aligns source motion and target motion temporally, contributing to high R-precision in editing tasks.

*Additional ablation studies are available in the supp. material.

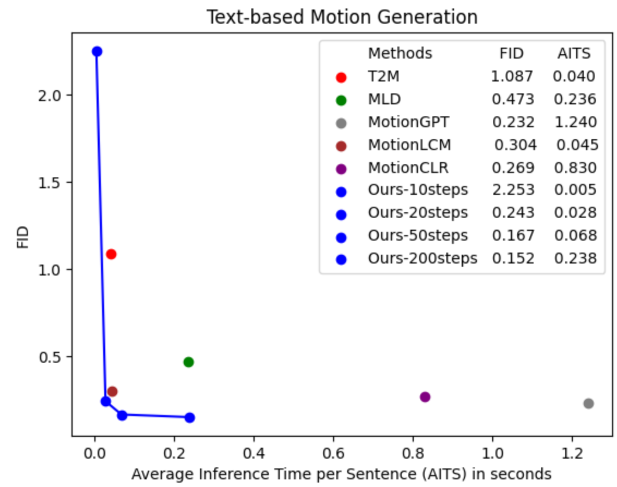


Figure 6. **Impact of timesteps during inference on MotionLab**. The closer to the lower left corner, the more powerful the model.

Additionally, we evaluate the impact of timesteps during inference on our MotionLab and compare its performance with baseline methods in terms of generation quality and inference time for the text-based motion generation task. As shown in Figure 6, our framework strikes an optimal balance between generation quality and efficiency.

7. Conclusion

Building on our proposed Motion-Condition-Motion paradigm, we have developed the MotionLab framework to unify human motion generation and editing. We have introduced the MotionFlow Transformer to leverage rectified flows to learn the mapping from source motion to target motion based on specified conditions. Additionally, we have incorporated Aligned Rotational Position Encoding to ensure synchronization between source motion and target motion, Task Instruction Modulation, and Motion Curriculum Learning for effective multi-task learning. Our proposed MotionLab framework demonstrates superior versatility, performance, and efficiency compared to existing state-of-the-art methods and our specialist models.

Acknowledgments. This research is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 1 – SMU-SUTD Internal Research Grant (SMU-SUTD 2023_02_09).

References

[1] 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):64–1, 2020. 1, 3

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

[3] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 3

[4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 1

[5] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9995, 2023. 1

[6] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 3, 6, 7

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 5

[8] 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. 2, 3, 6

[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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2, 3, 4, 6

[10] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 2

[11] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 1, 2, 3, 6

[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 *European Conference on Computer Vision*, pages 390–408. Springer, 2025. 1, 2, 3, 6

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 4

[16] Zhaoxin Fan, Longbin Ji, Pengxin Xu, Fan Shen, and Kai Chen. Everything2motion: Synchronizing diverse inputs via a unified framework for human motion synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1688–1697, 2024. 2, 3

[17] Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Flux that plays music. *arXiv preprint arXiv:2409.00587*, 2024. 3

[18] Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models. In *European Conference on Computer Vision*, pages 323–339. Springer, 2025. 1

[19] Purvi Goel, Kuan-Chieh Wang, C Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 2, 3

[20] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 3

[21] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 1, 6

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

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

[24] Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space. *arXiv preprint arXiv:2401.13505*, 2024. 1

[25] Ziyan Guo, Haoxuan Qu, Hossein Rahmani, Dewen Soh, Ping Hu, Qiuhong Ke, and Jun Liu. Tstmotion: Training-free scene-aware text-to-motion generation. *arXiv preprint arXiv:2505.01182*, 2025. 2

[26] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 2

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[29] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022. 1, 3

[30] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2, 3, 6

[31] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang Yu, and Jiayuan Fan. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, pages 54–74. Springer, 2025. 3

[32] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 3, 6

[33] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 3

[34] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. *arXiv preprint arXiv:2409.15904*, 2024. 3

[35] 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. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 3

[36] Zeyu Ling, Bo Han, Shiyang Li, Hongdeng Shen, Jikang Cheng, and Changqing Zou. Motionllama: A unified framework for motion synthesis and comprehension. *arXiv preprint arXiv:2411.17335*, 2024. 3

[37] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2, 3

[38] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2, 3

[39] 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. 1

[40] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 3

[41] Mingshuang Luo, Ruibing Hou, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *arXiv preprint arXiv:2405.16273*, 2024. 2, 3

[42] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024. 3

[43] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

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

[45] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 3

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

[47] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2024. 3

[48] Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *ACM Trans. Graph.*, 41(6):184–1, 2022. 3

[49] 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. 4

[50] Alessio Sampieri, Alessio Palma, Indro Spinelli, and Fabio Galasso. Length-aware motion synthesis via latent diffusion. *arXiv preprint arXiv:2407.11532*, 2024. 1

[51] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3, 6

[52] Aayam Shrestha, Pan Liu, German Ros, Kai Yuan, and Alan Fern. Generating physically realistic and directable human motions from multi-modal inputs. In *European Conference on Computer Vision*, pages 1–17. Springer, 2025. 2, 3

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[54] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498. PMLR, 2023. 1

[55] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, 2024. 1, 2, 3, 7

[56] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[57] Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. Lgtm: Local-to-global text-driven human motion diffusion model. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 1

[58] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 1, 2, 3

[59] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 6

[60] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv preprint arXiv:2406.00320*, 2024. 3

[61] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024. 3

[62] 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. 3

[63] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023. 1, 2, 3, 6, 7

[64] Han Yang, Kun Su, Yutong Zhang, Jiaben Chen, Kaizhi Qian, Gaowen Liu, and Chuang Gan. Unimumo: Unified text, music and motion generation. *arXiv preprint arXiv:2410.04534*, 2024. 2, 3

[65] 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 the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 3, 6

[66] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2, 3, 6

[67] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36:13981–13992, 2023. 1, 3

[68] 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 *European Conference on Computer Vision*, pages 397–421. Springer, 2025. 2, 3

[69] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, pages 405–421. Springer, 2025. 1, 2, 3

[70] Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5632–5641, 2023. 2, 3

[71] Zixiang Zhou, Yu Wan, and Baoyuan Wang. A unified framework for multimodal, multi-part human motion synthesis. *arXiv preprint arXiv:2311.16471*, 2023. 2, 3

[72] 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. 3