# FineMotion: A Dataset and Benchmark with both Spatial and Temporal Annotation for Fine-grained Motion Generation and Editing

Bizhu Wu[1,2,4] Jinheng Xie[5] Meidan Ding[1,4] Zhe Kong[6] Jianfeng Ren[2*] Ruibin Bai[2]
Rong Qu[7] Linlin Shen[2,3,4*]
[1] School of Computer Science & Software Engineering, Shenzhen University
[2] School of Computer Science, University of Nottingham Ningbo China, Ningbo, China
[3] Computer Vision Institute, School of Artificial Intelligence, Shenzhen University
[4] Guangdong Provincial Key Laboratory of Intelligent Information Processing
[5] National University of Singapore [6] Shenzhen Campus of Sun Yat-sen University
[7] School of Computer Science, University of Nottingham, Nottingham, United Kingdom
wubizhu@email.szu.edu.cn, jianfeng.ren@nottingham.edu.cn, llshen@szu.edu.cn

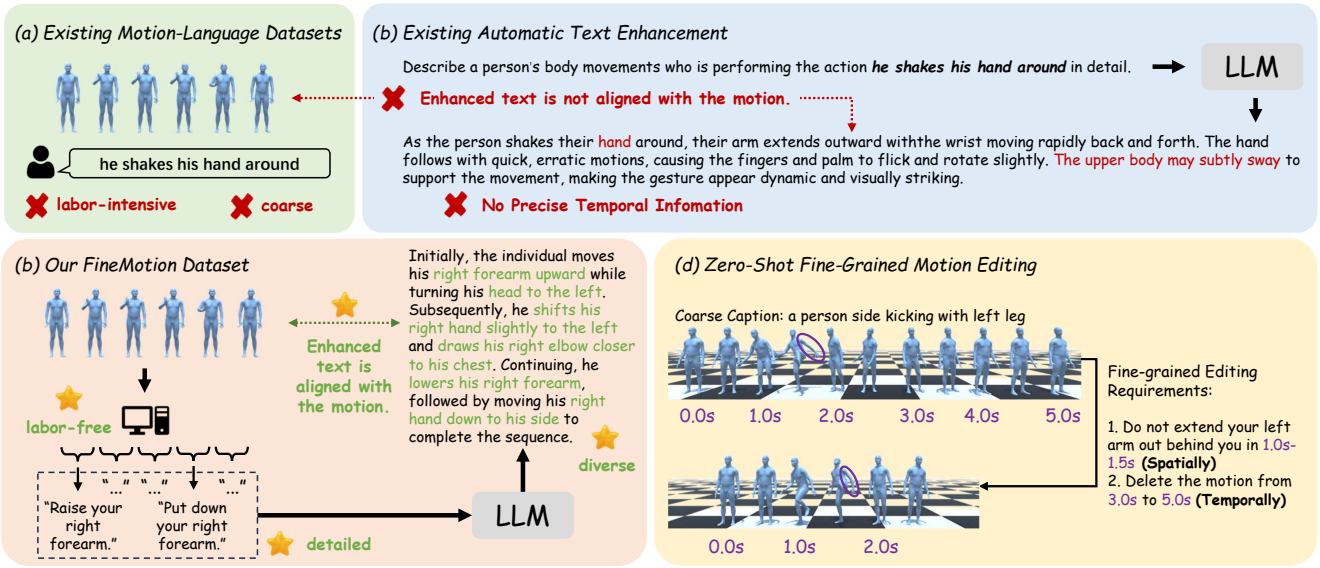



Figure 1. Illustration of **(a) Existing Motion-Language Datasets** are manually annotated, with textual descriptions that are coarse and lack detail. **(b) Existing textual enhancement** works obtained more detailed descriptions of a motion phrase or caption via large language models, but failed to align with the actual motion sequence. **(c) Our FineMotion dataset** features strictly aligned and fine-grained descriptions of human body part movements for both motion snippets (short segments of motion sequences) and entire motion sequences, while being easily scalable. **(d)** The proposed dataset further enables **zero-shot fine-grained motion editing** capabilities.

## Abstract

*Generating realistic human motions from textual descriptions has undergone significant advancements. However, existing methods often overlook specific body part movements and their timing. In this paper, we address this issue by enriching the textual description with more details. Specifically, we propose the FineMotion dataset, which contains over 442,000 human motion snippets — short segments of human motion sequences — and their corresponding detailed descriptions of human body part movements. Additionally, the dataset includes about 95k detailed paragraphs describing the movements of human body parts of entire motion sequences. Experimental results demonstrate the significance of our dataset on the text-driven fine-grained human motion generation task, especially with a remarkable +15.3% improvement in Top-3 accuracy for*

*Corresponding Author

*the MDM model. Notably, we further support a zero-shot pipeline of fine-grained motion editing, which focuses on detailed editing in both spatial and temporal dimensions via text. Dataset and code available at: CVI-SZU/FineMotion*

## 1. Introduction

The text-to-motion task involves generating human motion sequences from natural language textual descriptions, and has attracted growing interest due to its applications in animation making, virtual reality, and robotics. Recent advancements in this task [6, 12, 18, 25, 30] have advanced its practical deployment. Nowadays, people have higher expectations for this task, emphasizing the need to be controllable and realistic.

*"A man waves his right hand"* is an example textual description from the popular text-motion pair dataset, HumanML3D [6]. The motion sequences generated from this description may contain three stages: raising the right hand near the head, waving it, and lowering it. However, several questions arise: (1) *When should the right hand be raised or lowered?* (2) *How long should the wave last?* (3) *Are other body parts involved?* Obviously, textual descriptions in existing text-motion pair datasets, like HumanML3D [6] and KIT-ML [19], are **too coarse and less informative** to answer these questions.

To address these limitations, several works [1, 11] have proposed extending existing coarse captions with detailed descriptions using Large Language Models (LLMs). Since current LLMs cannot directly process motion sequences to generate corresponding detailed descriptions, these works rely on language-only LLMs. They prompt LLMs with only coarse captions as input to generate detailed body part movement descriptions based on LLMs' own biases, as illustrated in Fig. 1(b). However, it is evident that the enhanced textual output **fails to precisely align with the actual human motion sequences**.

In this work, we construct a new dataset, **FineMotion**, which provides precise Body Part Movement (BPM) descriptions for short temporal intervals in human motion sequences. It contains about 420k automatically generated BPM descriptions for motion snippets (short segments of motion sequences), referred to as **BPMSD**, and over 21k human-annotated ones. Additionally, it includes around 95k BPM Paragraph (**BPMP**) for detailed descriptions of entire human motion sequences. Examples are shown in Fig. 2. Notably, the temporal information embedded in the textual annotations allows for easy augmentation, such as random cropping along the temporal dimension to generate numerous pairs of motion clips (composed of several adjacent snippets) and their corresponding BPM descriptions.

Tab. 1 compares our FineMotion with existing text-motion pair datasets. The textual descriptions in the KIT-ML[19] and HumanML3D [6] datasets are coarse and lack detail. HuMMan-MoGen [32] provides detailed body part movement descriptions for each phase, but relies on manually specifying the start and end points of standardized phases, limiting scalability. In contrast, we not only include fine-grained textual descriptions for motions in our FineMotion dataset, but also propose an efficient and scalable pipeline for the automatic generation of detailed textual descriptions, facilitating easy dataset expansion.

With FineMotion, we establish a benchmark to evaluate several state-of-the-art motion generation methods. Comprehensive experiments demonstrate its effectiveness in producing precise and realistic motion. Building on this foundation, we further explore a zero-shot fine-grained motion editing pipeline, enabling users to modify descriptions to adjust motion content. This improves interaction efficiency and broadens applications.

Overall, our key contributions are as follows: **First**, we develop an efficient and scalable pipeline for the automatic generation of detailed motion descriptions. It offers a potent solution to the imprecision and lack of specificity issues in the motion description annotations domain. **Second**, the proposed FineMotion dataset bridges the domain gap with over 442k textual descriptions for short motion snippets, and around 95k paragraphs for whole motion sequences, all informative and strictly aligned. **Third**, we validate the dataset's effectiveness and generalization by benchmarking classical text-to-motion models that are intricately adapted and carefully tailored to handle our long and detailed text. Experimental results show that all these models exhibit notable performance gains, particularly with a +15.3% increase in Top-3 retrieval accuracy for our MDM variant. **Finally**, we demonstrate a zero-shot fine-grained motion editing pipeline, enabling controllable and realistic motion generation via textual modifications.

| Dataset | Year | Number of Motions | Number of Texts | Granularity | Annotation Source | Easily Scalable |
|---|---|---|---|---|---|---|
| KIT-ML [19] | 2016 | 3,911 | 6,278 | Coarse | Human | × |
| HumanML3D [6] | 2022 | 14,616 | 44,970 | Coarse | Human | × |
| HuMMan-MoGen [32] | 2023 | 2,968 | 102,336 | Fine | Human | × |
| **FineMotion (Ours)** | 2024 | 14,616 | 442,314 (Snippet) + 94,432 (Sequence) | Fine | Auto + Human | ✓ |

Table 1. Comparisons of 3D human motion-language datasets.

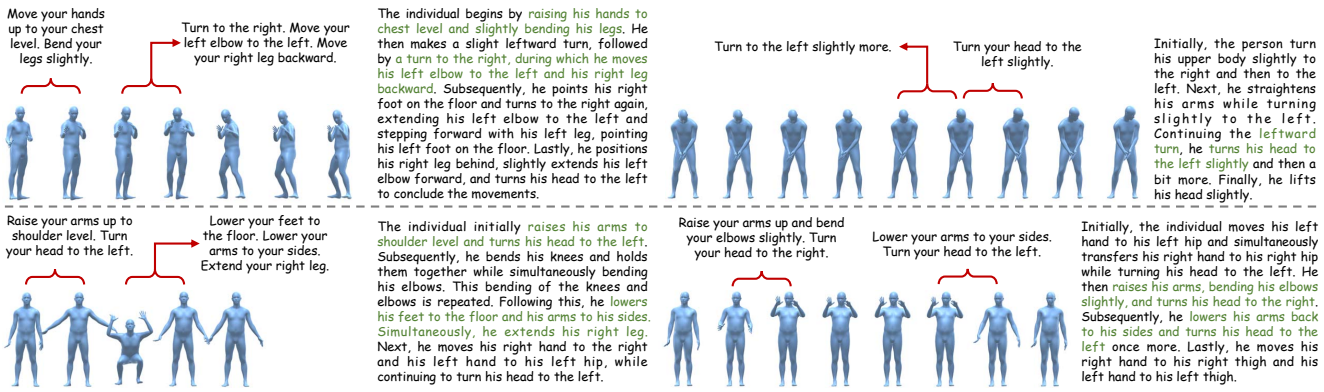



Figure 2. **Examples from the FineMotion dataset. Top**: Human-annotated BPM snippet descriptions and paragraphs. **Bottom**: Automatically generated BPM snippet descriptions and paragraphs. Colored text in paragraphs links to the corresponding snippet descriptions.

## 2. Related Work

**Text-Driven Human Motion Generation.** Motion generation can be generated from various conditions, including text [12, 18, 25, 30, 34], action classes [11, 17, 20, 24], and music [13, 14]. Among these, text-to-motion stands out for its user-friendly language interface. MotionDiffuse [31] and MDM [25] integrated diffusion models for text-to-motion generation with impressive results. T2M-GPT [30] and MoMask [8] quantized human motions and used transformer networks to generate high-quality motion. Recently, MotionGPT [10] treated motion as a foreign language in a natural language model. However, these methods rely on coarse captions that overlook fine-grained action details. In contrast, we focus on detailed textual descriptions that capture finer body part movements over time, helping generate motions more similar to the ground truth.

**Human Motion and Language Data.** Existing datasets like KIT-ML [19] and HumanML3D [6] offer textual annotations for 3D motions but lack fine-grained details. To address this, some works leveraged large language models for enhanced semantic annotations. Action-GPT [11] used carefully designed prompts to generate more detailed descriptions, but these may not align well with ground-truth motion. SemanticBoost [9] and MotionScript [29] mapped body part movements to predefined statuses but ignored precise temporal information and relied on fixed templates. FineMoGen [32] introduced the HuMMan-MoGen dataset with fine-grained spatio-temporal descriptions, but required extensive manual annotation of temporal boundaries, limiting scalability. Building on prior work, we present FineMotion, a new dataset with detailed, temporally precise, diverse, and motion-aligned annotations. Besides, our automatic dataset construction pipeline facilitates easy scaling.

**Text-Driven Human Motion Editing.** Several text-to-motion approaches [12, 24–26] have explored human motion editing. Diffusion model-based methods [12, 25, 26] diffused a reference motion, masked specific frames and joints, replaced the masked parts with the ones conditioned on another coarse text, and denoised to obtain the edited human motion. MotionCLIP [24] performs editing via latent space arithmetic. However, these methods rely on coarse descriptions, which limit their control over specific body parts, timing, and duration. In contrast, our baseline models generate motion strictly based on detailed descriptions, facilitating fine-grained motion editing across temporal and spatial dimensions through precise text editing.

## 3. The proposed FineMotion dataset

The proposed FineMotion dataset builds upon HumanML3D [6] by describing motions in fine details both spatially and temporally. The motion sequences, sourced from AMASS [16] and HumanAct12 [5], span diverse actions like 'walking', 'swimming', and 'dancing'. They are pre-processed by scaling to 20 FPS, randomly cropping those longer than 10 seconds, re-targeting to a standard skeletal template, and rotating to face the Z+ direction.

As for the textual descriptions, we include two types: One is the body part movement description for the motion snippet, a short segment from the motion sequence, short for **BPMSD**; The other one is the body part movement description paragraph for the whole motion sequence, short for **BPMP**. Generally, the enriched textual descriptions have the following three properties: (1) *more fine-grained descriptions of body part movements*, (2) *precise temporal information*, and (3) *more diverse*. We next present the dataset construction pipeline and some dataset statistics.

### 3.1. Dataset Construction Pipeline

The pipeline for enhancing text descriptions from a human motion sequence is illustrated in Fig. 3. The input is the SMPL [15] pose parameters of the human motion sequence, while the pipeline outputs two types of BPM descriptions over time in English. *i.e.*, BPM Snippet Description and BPM Paragraph. The process comprises three steps:

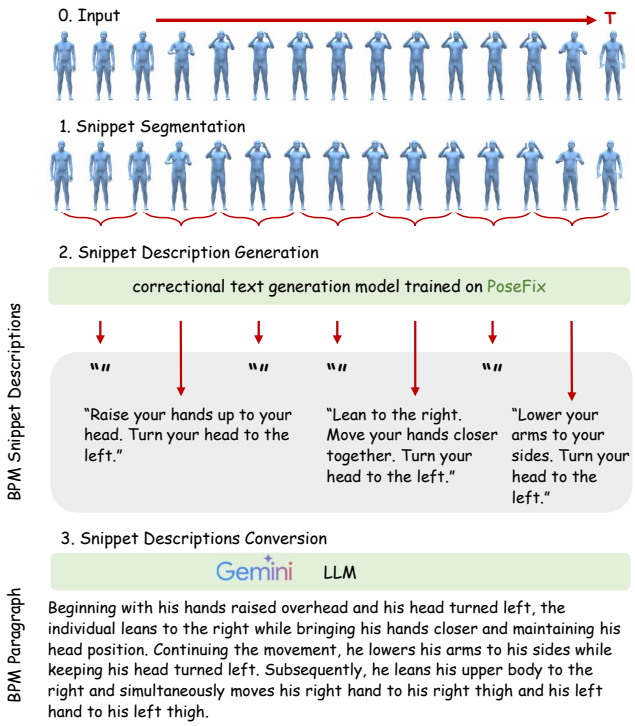



Figure 3. The construction pipeline of our FineMotion dataset.

1. Given a motion sequence, we divide it into short snippets along the temporal dimension (Sec. 3.1.1).
2. The detailed BPM descriptions are generated for each snippet (Sec. 3.1.2).
3. All snippet descriptions from the same motion sequence are further organized into a paragraph (Sec. 3.1.3).

Notably, this pipeline is **universal** and can be applied to any text-motion pair datasets or motion-only datasets.

### 3.1.1. Snippet Segmentation

To obtain detailed, temporally aligned descriptions of motions, we first explicitly segment each motion into short snippets along the temporal dimension.

In this dataset, we choose to fix the snippet duration for two main reasons: First, a fixed snippet duration simplifies the dataset scaling process within our automated dataset construction pipeline. It **minimizes the need for manual annotation to determine the start and end points of each snippet**. Secondly, a consistent duration reduces the complexity of the fine-grained motion generation model, as it **eliminates the need for additional inputs, such as the start and end points for each snippet**. Otherwise, the model would require these inputs to be explicitly aligned with each snippet's fine-grained description.

To determine the optimal snippet duration $T_s$, we propose two guiding principles to help researchers tailor this value to their own datasets: First, **select $T_s$ to minimize similarity between snippets' start and end poses**. As shown in Fig. 4, we calculate cosine similarity between

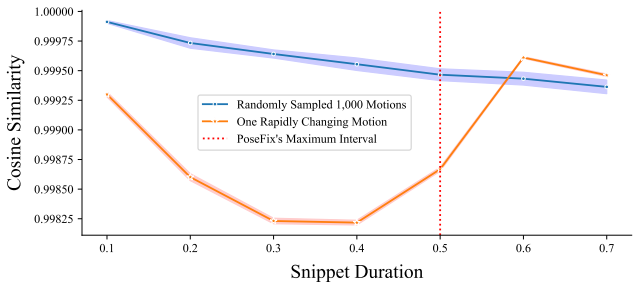



Figure 4. Mean and 95% confidence interval of the cosine similarity between semantic features of snippets' start and end poses.

PoseScript [2] semantic features of snippets' start and end poses. The analysis was performed on randomly sampled snippets from all motions in our dataset (*blue* line) and a representative example of a rapidly changing motion (*orange* line), with various durations. Results show that motions in our dataset generally progress slowly, suggesting that a longer interval helps reduce redundancy among snippet descriptions. Further details are available in the appendix. Meanwhile, PoseFix [3] suggests that larger time differences between two poses allow for a wide range of plausible in-between motions. Therefore, the second principle is that **the value of $T_s$ should not exceed 0.5s**, which is the maximum time difference for pose pair selection specified by PoseFix [3]. Following these principles, we set $T_s$ to 0.5s. Notably, any remaining segment of a motion sequence shorter than $T_s$ is also treated as an individual snippet.

### 3.1.2. Snippet Description Generation

**Collection of Automatically Generated Annotations.** This step builds upon the outstanding foundation work, PoseFix [3]. It developed a correctional text generation model to describe how body parts in a source pose should be modified to achieve a target pose. Specifically, the model integrates two pose embeddings in the cross-attention mechanisms of a text transformer to generate correctional text for a given pose pair. We believe that the correctional text describing the transition between start and end poses of a snippet effectively captures the body part movements within it, and can thus be naturally regarded as the detailed BPM description for this snippet.

**Collection of Human Annotations.** We recruited eight undergraduate students with strong English comprehension skills to manually annotate motion sequences. Specifically, we first constructed a BPM description corpus, which is composed of sentences derived from the automatically generated annotations for all snippets. Then, for each motion sequence, annotators were provided with its automatically generated annotations to ease their workload. If those annotations were inappropriate, they were instructed to select suitable sentences from the corpus. This ensured that the manually annotated descriptions closely match the style of

the automatically generated ones. Besides, annotators were required to remove trivial or passive body part movements (such as '*move your left leg backward*' in a '*walking forward*' motion sequence) from each snippet description. For snippets involving rapid changes (*i.e.*, containing multiple stages), annotators were encouraged to describe each stage separately and use phrases such as "*Then,*" to connect them.

After the preliminary annotations, two additional well-trained students conducted two rounds of manual checks to scrutinize, correct, and supplement the content. To fix any remaining grammar and spelling issues, we utilized Grammarly [4]. Overall, we annotated around 21k motion snippets (*i.e.*, 5% of the motion sequences). The upper part of Fig. 2 shows examples of motion sequences with their manually annotated BPM snippet descriptions; more can be found in the appendix.

#### 3.1.3. Paragraph Generation

Apart from the detailed BPM description for the motion snippet, we also include that for the whole motion sequence in the FineMotion dataset, referred to as the BPM paragraph. Instead of applying fixed templates to connect the descriptions for all the snippets in a single human motion sequence, we leverage Gemini's [23] advanced language capabilities to create diverse, coherent paragraphs that follow natural human logic.

Specifically, for each human motion sequence, we first remove empty BPM snippet descriptions, and connect the rest with numbers to preserve temporal order, resulting in *BPMSDs* (see example in ### Input ###). Following ActionGPT [11], we craft a prompt to guide Gemini to organize these descriptions. After multiple trials, we determine the following prompt. It first introduces the task, *i.e.*, arranging all snippet descriptions into a cohesive paragraph. Then, it lists several requirements to ensure continuity, proper time order, converting PoseFix's imperative output to descriptive ones, reliability, and completeness for the output paragraph. An example is also included to help Gemini understand and perform the conversion effectively.

> Given a list of items describe the body part's movements that happen in succession. Note that each item describes the body part's movements that happen simultaneously. The items are already arranged in chronological order. Please organize these items into a single paragraph according to the following requirements:
>
> (1) If some of the body part's movements in adjacent two items are similar, then use vocabulary indicating their continuity, such as 'keep' and 'continue'.
> (2) Use proper conjunctions of time to connect all items but do not add too many additional descriptions.
> (3) Change the output into descriptive statements. For example, the subject can be changed to 'The person', 'The individual', etc.
> (4) Make sure not to add any extra body part's movement that does not exist in the original list of items.
> (5) Make sure all the body part's movements in the list are described in the output.
>
> ### Input ###: 1. Lower your left leg, keeping the point at your foot and curving it very slightly. Bring down your right hand even to your left hand. 2. Raise your right leg. Bring down your right hand. 3. Lower your legs so your feet and upper legs are parallel with the ground. Bend your right arm slightly to have it hover over your elevated legs. Lower your left forearm so it is at an angle pointing down.
>
> ### Output ###: The individual lowers his left leg, ensuring the foot is pointed and slightly curved, while simultaneously bringing his right hand down to meet the left. Afterward, he raises his right leg and continues bringing the right hand down. Consequently, he lowers his legs until the feet and upper legs are parallel to the floor. At the same time, he slightly bends his right arm, positioning it above his elevated legs, and lowers his left forearm, forming a downward angle.
>
> ### Input ###: [*BPMSDs*]
>
> ### Output ###:

One should notice that the input to Gemini here is all BPM snippet descriptions within a motion sequence. All Gemini has to do is connect these precise snippet descriptions into a coherent paragraph. As a result, the generated BPM paragraphs are **strictly aligned with the actual motion sequence**, which addresses the limitations mentioned in Sec. 1. Interestingly, instead of merely connecting all the snippet descriptions, Gemini sometimes offers precise paraphrases, further **enhancing the diversity** of BPM paragraphs. Examples can refer to Fig. 2 and the appendix.

### 3.2. Statistics Analysis

In summary, our FineMotion dataset contains 21,346 human-annotated BPM descriptions and 420,968 automatically generated ones for diverse motion snippets, *i.e.*, **BPMSD**. Notably, the temporal information within the textual annotations facilitates **easy augmentation**, such as performing random cropping along the temporal dimension. This approach can generate numerous pairs of motion clips—each consisting of several adjacent snippets—along with their corresponding BPM descriptions. Additionally, the dataset includes 4,492 BPM paragraphs (**BPMP**) organized from human-annotated BPM snippet descriptions, and 89,940 paragraphs organized from automatically generated BPM snippet descriptions, covering a total of 29,232 motion sequences.

For BPMSD, the average and median lengths are 18 and 19 words, respectively. As for BPMP, the average and median lengths are 247 and 221 words, respectively. In terms of **data splits**, we adopt the same ones as HumanML3D, in which the data is partitioned into 80%, 5%, and 15% for training, validation, and testing, respectively, ensuring no overlap among the subsets. Fig. 5 shows the most frequent words in the FineMotion dataset. Thanks to the assistance of Gemini, the vocabulary in BPMP (the second and fourth instances) is notably more diverse.


Figure 5. **Visualizations of the 200 most frequent words in our textual descriptions.** From left to right are the word clouds of automatically-generated BPMSD, automatically-generated BPMP, human-annotated BPMSD, and human-annotated BPMP.

## 4. Experiments

In this section, we first validate the accuracy of our textual annotation pipeline. Then, we build a text-driven fine-grained human motion generation benchmark on FineMotion. Finally, we explain how we have implemented zero-shot fine-grained motion editing with our dataset.

### 4.1. Data Preprocessing

We follow [6] to pre-process motion sequences into $d = 263$-dimensional features. For textual input, we use coarse descriptions in HumanML3D [6] (denoted as '**T**'), and the detailed BPM snippet descriptions (BPMSDs) or BPM paragraph (BPMP) from our FineMotion dataset (denoted as '**DT**'). Notably, since a BPMSD covers only a short interval rather than the entire sequence, we use a fixed template to connect all BPMSDs in a motion sequence. Specifically, empty snippet descriptions, which indicate no significant BPM, are replaced with the special token <Motionless>. We then use the special token <SEP> to connect snippet descriptions across intervals and preserve temporal information, *e.g.*,

Given all the BPM snippet descriptions from a motion sequence:
["", "", "Move your right leg forward slightly.", "Turn to the left. Move your left leg forward. Move your left hand back slightly.", "Lean to the right. Move your right leg forward."]

Fit into the template:
"<Motionless> <SEP> <Motionless> <SEP> Move your right leg forward slightly. <SEP> Turn to the left. Move your left leg forward. Move your left hand back slightly. <SEP> Lean to the right. Move your right leg forward."

### 4.2. Evaluation Metrics

We evaluate the generated motions using metrics from [6, 30]: Frechet Inception Distance (FID), Multi-modal Distance (MM-Dist), R-Precision Top-1/2/3, Diversity, and Multi-modality (MModality). These metrics assess the realism and diversity of synthesized motions, with definitions provided in [6, 30].

### 4.3. Baseline Models

To accommodate both coarse and detailed descriptions, we intricately adapted three classical text-to-motion models, *i.e.*, MDM [25], T2M-GPT [30], and MoMask [8], to better handle our long, detailed text. Specifically, we replace the CLIP [21] text encoder with T5-Base [22] to avoid truncation of over-length detailed text. We then apply mean pooling along the sequence length dimension of the T5-Base encoder output to obtain a single text embedding. Notably, instead of directly connecting coarse captions and detailed text into one before encoding, our variants are trained to synthesize motion from the concatenated embeddings of both components. We refer to these adapted variants as (T&DT)2M-MDM, (T&DT)2M-GPT, and (T&DT)2M-MoMask, respectively. More details on model design and ablation studies can be found in the appendix.

### 4.4. Evaluation of the Textual Annotation Pipeline

We validate the quality of our detailed text using our (T&DT)2M-GPT variant in Tab. 2, where all models are trained under the same setting. For fair comparisons, we re-implement T2M-GPT with a T5-Base text encoder as the *baseline* for the coarse-grained text-to-motion (T2M) task. Baseline results are reported in Tab. 2.(1). Additional results for other variants are provided in the appendix.

**Automatically Generated Annotations.** We train (T&DT)2M-GPT on both the T2M task and fine-grained text-to-motion task, denoted as (T&DT)2M. Here, DT is automatically generated descriptions (**DT$^{\text{Auto}}$**). When training the T2M task, DT is replaced by the special token <EMPTY>. From Tab. 2.(2), (T&DT)2M-GPT achieve performance with an FID of 0.091 (*vs.* 0.123 of *baseline*) and R-Top3 of 0.789 (*vs.* 0.781 of *baseline*). Similar gains are observed with our automatically generated BPMP in

| | Train Set | | | Test Set | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | T2M | | (T&DT$^{\text{Auto}}$)2M | | (T&DT$^{\text{Human}}$)2M | |
| | T2M | (T&DT$^{\text{Auto}}$)2M | (T&DT$^{\text{Human}}$)2M | R-Top3 ↑ | FID ↓ | R-Top3 ↑ | FID ↓ | R-Top3 ↑ | FID ↓ |
| (1) | ✓ | - | - | $0.781^{\pm.002}$ | $0.123^{\pm.005}$ | - | - | - | - |
| *DT: BPMSD* | | | | | | | | | |
| (2) | ✓ | ✓ | - | $0.784^{\pm.002}$ | $0.124^{\pm.005}$ | $\mathbf{0.789}^{\pm.002}$ | $\mathbf{0.091}^{\pm.003}$ | - | - |
| (3) | ✓ | ✓ | ✓ | $0.781^{\pm.002}$ | $0.154^{\pm.007}$ | $\mathbf{0.789}^{\pm.002}$ | $0.112^{\pm.005}$ | $\mathbf{0.789}^{\pm.002}$ | $\mathbf{0.091}^{\pm.004}$ |
| *DT: BPMP* | | | | | | | | | |
| (4) | ✓ | ✓ | - | $0.779^{\pm.002}$ | $0.136^{\pm.005}$ | $0.785^{\pm.002}$ | $0.102^{\pm.004}$ | - | - |
| (5) | ✓ | ✓ | ✓ | $0.781^{\pm.002}$ | $0.155^{\pm.006}$ | $\mathbf{0.788}^{\pm.002}$ | $0.104^{\pm.005}$ | $\mathbf{0.788}^{\pm.002}$ | $\mathbf{0.100}^{\pm.005}$ |

Table 2. **Evaluation of our textual annotation pipeline with (T&DT)2M-GPT.** 'T' means coarse descriptions on the HumanML3D, while 'DT' means detailed texts on our FineMotion dataset. We repeat all evaluations 20 times and report the average with a 95% confidence interval. **Bold** text means the best results in each block. Results show that incorporating our fine-grained and human-annotated texts enhances motion generation performance, which proves the quality of our textual annotation pipeline.

| Methods | Text Granularity | | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → | MModality ↑ |
|---|---|---|---|---|---|---|---|---|---|
| | Coarse (T) | Fine (DT) | Top-1 | Top-2 | Top-3 | | | | |
| Real motion | ✓ | - | $0.511^{\pm.003}$ | $0.703^{\pm.003}$ | $0.797^{\pm.002}$ | $0.002^{\pm.000}$ | $2.974^{\pm.008}$ | $9.503^{\pm.065}$ | - |
| TEMOS [18] ECCV'22 | ✓ | - | $0.424^{\pm.002}$ | $0.612^{\pm.002}$ | $0.722^{\pm.002}$ | $3.734^{\pm.028}$ | $3.703^{\pm.008}$ | $8.973^{\pm.071}$ | $0.368^{\pm.018}$ |
| TM2T [7] ECCV'22 | ✓ | - | $0.424^{\pm.003}$ | $0.618^{\pm.003}$ | $0.729^{\pm.002}$ | $1.501^{\pm.017}$ | $3.467^{\pm.011}$ | $8.589^{\pm.076}$ | $2.424^{\pm.093}$ |
| Guo et al.[6] CVPR'22 | ✓ | - | $0.455^{\pm.003}$ | $0.636^{\pm.003}$ | $0.736^{\pm.002}$ | $1.087^{\pm.021}$ | $3.347^{\pm.008}$ | $9.175^{\pm.083}$ | $2.219^{\pm.074}$ |
| MotionDiffuse [31] TPAMI'24 | -✓ | - | $0.491^{\pm.001}$ | $0.681^{\pm.001}$ | $0.782^{\pm.001}$ | $0.630^{\pm.001}$ | $3.113^{\pm.001}$ | $9.410^{\pm.049}$ | $1.553^{\pm.042}$ |
| Fg-T2M [28] ICCV'23 | ✓ | - | $0.492^{\pm.002}$ | $0.683^{\pm.003}$ | $0.783^{\pm.002}$ | $0.243^{\pm.019}$ | $3.109^{\pm.007}$ | $9.278^{\pm.072}$ | $1.614^{\pm.049}$ |
| FineMoGen [28] NeurIPS'23 | ✓ | - | $0.504^{\pm.002}$ | $0.690^{\pm.002}$ | $0.784^{\pm.002}$ | $0.151^{\pm.008}$ | $2.998^{\pm.008}$ | $9.263^{\pm.094}$ | $2.696^{\pm.079}$ |
| MDM [25] arXiv'22 † | ✓ | - | $0.323^{\pm.006}$ | $0.498^{\pm.007}$ | $0.606^{\pm.008}$ | $3.137^{\pm.183}$ | $4.373^{\pm.043}$ | $9.525^{\pm.086}$ | $2.614^{\pm.102}$ |
| **(T&DT)-MDM** (BPMSD) | ✓ | ✓ | $0.445^{\pm.007}$ | $0.640^{\pm.009}$ | $0.745^{\pm.008}$ | $0.756^{\pm.081}$ | $3.412^{\pm.030}$ | $9.640^{\pm.095}$ | $2.495^{\pm.053}$ |
| **(T&DT)-MDM** (BPMP) | ✓ | ✓ | $0.460^{\pm.005}$ | $0.655^{\pm.005}$ | $0.759^{\pm.005}$ | $0.488^{\pm.046}$ | $3.276^{\pm.021}$ | $9.869^{\pm.108}$ | $2.340^{\pm.054}$ |
| T2M-GPT [30] CVPR'23 † | ✓ | - | $0.499^{\pm.003}$ | $0.688^{\pm.003}$ | $0.781^{\pm.002}$ | $0.123^{\pm.005}$ | $3.076^{\pm.009}$ | $9.747^{\pm.093}$ | $1.890^{\pm.085}$ |
| **(T&DT)2M-GPT** (BPMSD) | ✓ | ✓ | $0.510^{\pm.002}$ | $0.695^{\pm.002}$ | $0.789^{\pm.002}$ | $0.091^{\pm.004}$ | $3.002^{\pm.008}$ | $9.592^{\pm.079}$ | $1.594^{\pm.075}$ |
| **(T&DT)2M-GPT** (BPMP) | ✓ | ✓ | $0.506^{\pm.002}$ | $0.694^{\pm.002}$ | $0.788^{\pm.002}$ | $0.100^{\pm.005}$ | $3.023^{\pm.010}$ | $9.602^{\pm.057}$ | $1.615^{\pm.016}$ |
| MoMask [8] CVPR'24 † | ✓ | - | $0.466^{\pm.003}$ | $0.655^{\pm.003}$ | $0.753^{\pm.002}$ | $0.249^{\pm.012}$ | $3.359^{\pm.008}$ | $9.676^{\pm.083}$ | $1.371^{\pm.048}$ |
| **(T&DT)-MoMask** (BPMSD) | ✓ | ✓ | $0.519^{\pm.002}$ | $0.715^{\pm.002}$ | $0.811^{\pm.001}$ | $0.088^{\pm.003}$ | $2.946^{\pm.005}$ | $9.702^{\pm.075}$ | $1.271^{\pm.030}$ |
| **(T&DT)-MoMask** (BPMP) | ✓ | ✓ | $0.520^{\pm.003}$ | $0.717^{\pm.002}$ | $0.813^{\pm.002}$ | $0.055^{\pm.002}$ | $2.935^{\pm.009}$ | $9.679^{\pm.085}$ | $1.281^{\pm.051}$ |

Table 3. **Benchmark of FineMotion & Comparisons with HumanML3D.** We conduct all evaluations 20 times, reporting the average with a 95% confidence interval, except for MModality, which is run 5 times. '→' means results are better if the metric is closer to the real motions. For methods marked with †, we re-implement them using the same text encoder (T5) as ours to ensure fair comparisons. All our variants exhibit performance improvements, with (T&DT)-MDM showing a notable +15.3% increase in Top-3 retrieval accuracy.

Tab.2.(4). These results demonstrate that FineMotion's detailed BPM texts help generate motions more aligned with ground-truth. Furthermore, the BERTScore between human annotations and automatically generated ones is 0.89, comparable to the scores achieved by translation models such as Transformer-big on WMT14 En-De (0.86) and En-Fr (0.89) [33]. These pieces of evidence prove the effectiveness and quality of our automatically generated annotations.

**Human Annotations.** Similarly, we train (T&DT)2M-GPT on both the T2M and (T&DT)2M tasks, using human annotations as DT ('$\mathbf{DT^{Human}}$') when available and automatically generated ones otherwise. From Tab. 2.(3) and (5), human annotations offer more precise guidance, and further enhance performance, achieving an FID of 0.091 and an R-Top3 of 0.789 for BPMSD, and an FID of 0.100 and an R-Top3 of 0.788 for BPMP. In the following experiments, we adopt the training setting from Tab. 2.(3) and (5), and use (T&$\text{DT}^{\text{Human}}$)2M as the default test configuration.

## 4.5. Impact on Text-driven Motion Generation

**Benchmarking FineMotion.** From Tab. 3, including our fine-grained texts improves all variants. (T&DT)-MoMask achieves the best overall performance but the lowest MModality score, indicating reduced motion diversity. (T&DT)2M-GPT performs competitively while preserving high MModality, demonstrating our dataset's potential to enhance GPT-based methods. (T&DT)-MDM attains the highest MModality but the lowest R-Precision, suggesting it generates noisy and jittery motions.

**Comparison with HumanML3D.** To validate the significance of our dataset, we conduct a comparative analysis between FineMotion and HumanML3D (Coarse Text Only, *i.e.*, 'T') in Tab. 3. We re-implement MDM [25], T2M-GPT [30], and MoMask [8] on HumanML3D, replacing their text encoders with T5-Base [22] for fair comparisons. Overall, all variants trained on FineMotion consistently outperform those using HumanML3D's coarse descriptions (†), achieving better FID and R-Precision. Notably, MDM improves Top-3 retrieval accuracy by +15.3%, while (T&DT)-MoMask achieves the best FID and R-Precision across both datasets. For the suboptimal Diversity and MModality of our variants, we attribute this to the fine-grained descriptions, which constrain motion variation. These results underscore the versatility of FineMotion and its potential for zero-shot fine-grained motion editing.

## 4.6. Zero-shot Fine-Grained Motion Editing

Existing motion editing works rely on coarse captions that lack detail, limiting control over body part movements, timing, and duration. To address this, we use (T&DT)2M-GPT as an example and train it with pairs of coarse captions, temporally augmented motions, and corresponding detailed texts. The augmented data enhances the model's ability to align motions with BPM snippet descriptions. Thus, (T&DT)2M-GPT model can precisely control body part actions at specific time intervals based on detailed texts.

Based on this, we achieve the effect of fine-grained motion editing by the zero-shot pipeline in illustrated Fig. 6. It begins with users providing a text-to-motion model with a coarse description to synthesize an initial motion. Detailed descriptions are then obtained following the dataset con-

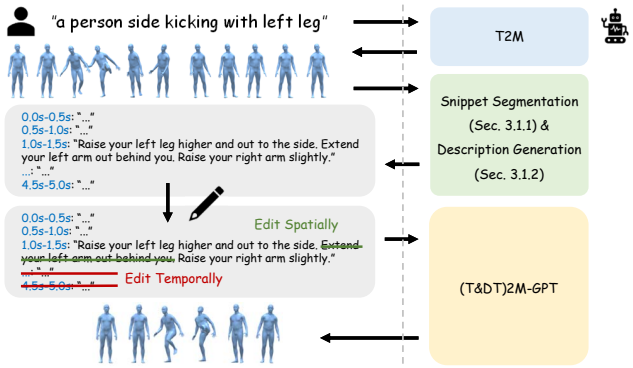



Figure 6. **Pipeline for zero-shot fine-grained motion editing.** To edit human motion with fine granularity, users first provide a coarse textual description of the desired motion. An initial motion is generated using any text-to-motion (T2M) model. This motion is then processed through the dataset construction pipeline to extract its BPM snippet descriptions. Users refine these descriptions with detailed editing instructions. Finally, the baseline model generates the fine-grained edited motion by adhering to both the modified BPM snippets and the original coarse caption.

struction pipeline, allowing users to refine them with fine-grained editing requirements. Finally, (T&DT)2M-GPT generates a new motion sequence from scratch based on the modified description, producing the final edited result. Notably, the re-generation process may introduce unintended changes beyond the specified regions.

Since quantitative metrics for evaluating editing results are unavailable, we follow [12] and [25] by conducting a user study with 30 randomly selected participants. Each user ranked 9 cases across 3 perspectives, totaling 27 questions. We compare our editing pipeline against T2M-GPT [30], a generative-based motion generation model, and FLAME [12], a diffusion-based text-driven motion editing approach. As these models lack fine-grained textual training data, their editing results are generated using coarse cap-

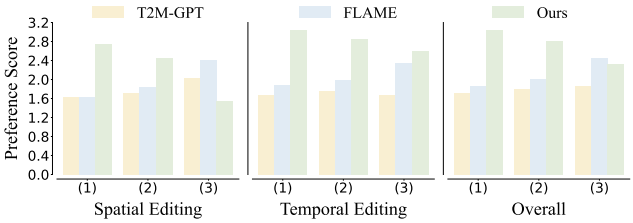



Figure 7. **The statistical results of the user study.** The *left* figure displays the average preference score for the spatial editing results (3 cases) of three models, with a score of 3 for the best model, 2 for the second, and 1 for the last. The *middle* one shows the score for the temporal editing results (6 cases). The *right* one summarizes the results for all 9 cases. Each case is evaluated from three perspectives: (1) whether the edited motion meets the editing requirements, (2) the naturalness of the edited motion, and (3) the similarity between the edited motion and the original one.

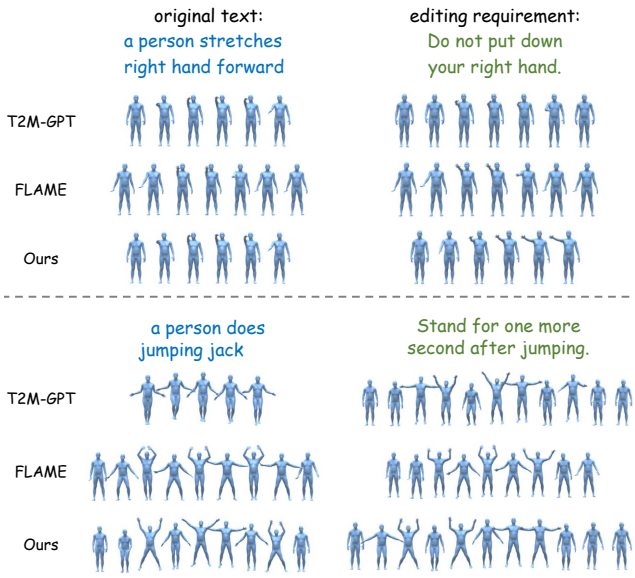



Figure 8. Examples of editing results where our pipeline achieves the highest preference score in terms of *meeting the editing requirements*. **Top:** Spatial Editing. **Bottom:** Temporal Editing.

tions describing the post-edited motions.

As shown in Fig. 7, users clearly preferred the edited results from our pipeline, as the other two methods often failed to meet the editing requirements. Our edited motions were also rated as the most natural. In terms of similarity to the original motions, our pipeline ranked second, as the other methods frequently produced identical outputs (*i.e.*, higher similarity) due to their inability to perform edits effectively. Notably, our pipeline excels in temporal editing, particularly in adjusting motion length, a task the other methods struggled with. This advantage stems from the temporally augmented data used in training (T&DT)2M-GPT, making temporal editing easier than spatial editing. Fig. 8 presents examples where our pipeline's edits were clearly favored by users compared to other methods. The user study is presented in the appendix.

## 5. Conclusion

This paper introduces FineMotion, a comprehensive dataset featuring human motion sequences paired with BPM snippet descriptions and paragraphs. We also develop an automated annotation pipeline to enable efficient dataset scaling. To validate the dataset's significance, we adapt three classical text-to-motion methods and benchmark them using our detailed textual annotations. Experimental results demonstrate that our detailed text improves motion generation performance, paving the way for zero-shot fine-grained motion editing. A user study confirms the high quality of our editing results. We anticipate that this exploratory work will shed light on future research in developing effective fine-grained motion understanding systems.

## Acknowledgements

This work was supported in part by National Key R&D Program of China (No. 2024YFF0618400), National Natural Science Foundation of China under Grant 82261138629, Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030, Guangdong Provincial Key Laboratory under Grant 2023B1212060076, and the Ningbo Municipal Bureau of Science and Technology under Grant 2023Z138, 2023Z237, 2024Z110 and 2024Z124.

## References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation supplementary material. 2

[2] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022. 4, 1

[3] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: Correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15018–15028, 2023. 4, 1

[4] Grammarly. https://www.grammarly.com. 5

[5] 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

[6] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 3, 6, 7

[7] 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. 7

[8] 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, 6, 7, 4

[9] Xin He, Shaoli Huang, Xiaohang Zhan, Chao Wen, and Ying Shan. Semanticboost: Elevating motion generation with augmented textual cues. *arXiv preprint arXiv:2310.20323*, 2023. 3

[10] 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, 2024. 3

[11] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 31–36. IEEE, 2023. 2, 3, 5

[12] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 2, 3, 8

[13] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1279, 2022. 3

[14] Ronghui Li, Yuqin Dai, Yachao Zhang, Jun Li, Jian Yang, Jie Guo, and Xiu Li. Exploring multi-modal control in music-driven dance generation. *arXiv preprint arXiv:2401.01382*, 2024. 3

[15] 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. 3

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

[17] 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

[18] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2, 3, 7

[19] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 3

[20] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13873–13883, 2023. 3

[21] 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. 6, 3

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6, 7, 3, 4

[23] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5

[24] 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. 3

[25] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3, 6, 7, 8, 4

[26] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 3

[27] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[28] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 7

[29] Payam Jome Yazdian, Eric Liu, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634*, 2023. 3

[30] 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. 2, 3, 6, 7, 8, 4

[31] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3, 7

[32] 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, 2024. 2, 3

[33] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 7

[34] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 3

# FineMotion: A Dataset and Benchmark with both Spatial and Temporal Annotation for Fine-grained Motion Generation and Editing

Supplementary Material

## 6. License

The license for human motion sequences in this dataset follows the term specified at HumanML3D and AMASS. The textual descriptions in our FineMotion dataset are under the CC BY 4.0 International license. For detailed license information, please refer to https://creativecommons.org/licenses/by/4.0/legalcode

## 7. Discussion on Selecting of optimal $T_s$

To determine the optimal snippet duration $T_s$, we propose two guiding principles to help researchers tailor this value to their own datasets. As shown in Fig. 4, we randomly sample 1,000 snippets with varying durations from all motion sequences in our dataset. Then, we calculate the cosine similarity between the PoseScript [2] semantic features of the start and end poses for each snippet. A higher cosine similarity indicates that the start and end poses are more similar, suggesting that the motion progresses slowly; conversely, a lower similarity indicates faster progression.

Our results show that the motions in our dataset generally progress slowly, prompting the selection of a larger interval to avoid redundancy. Here, we also display the statistical results of a rapidly changing motion (1,000 random samples of start and end points) in Fig. 4. The results indicate that the similarity of the pose semantic features first decreases and then increases as the temporal interval grows. From this, we derive the first principle for selecting the optimal value of $T_s$: **Choose the value of $T_s$ that minimizes the similarity between the start and the end poses.** Meanwhile, PoseFix [3] suggests that larger time differences between two poses allow for a wide range of plausible in-between motions. Therefore, the second principle is that **the value of $T_s$ should not exceed 0.5s**, which is the maximum time difference for pose pair selection specified by PoseFix [3]. Following these principles, we set $T_s$ to 0.5s. Notably, any remaining segment of a motion sequence shorter than $T_s$ is also treated as an individual snippet.

## 8. Data Format Examples

The data format example for all the detailed human body part snippet descriptions (BPMSDs) in a whole human motion sequence is shown below:

```
{
    "000314":              # name of motion sequence
    [
        "",                # 0.0s-0.5s' BPMSD
        "Bend your elbows and raise your hands up to your head.",
        "",                # 1.0s-1.5s' BPMSD
        "",                # 1.5s-2.0s' BPMSD
        "Turn your upper body to the right slightly.",
        "",                # 2.5s-3.0s' BPMSD
        "Straighten your elbows and lower your hands to your thighs.",
        "Straighten your elbows completely and move your hands back to your
        sides.",
    ],
}
```

The data format example for three different detailed human body part paragraphs (BPMPs) for the same human motion sequence is shown below:

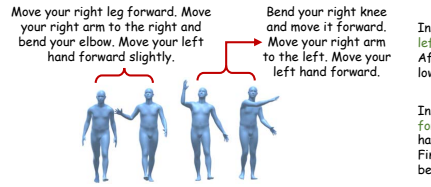
```
{
    "000314":              # name of motion sequence
    [
        "Initially, the person bends his elbows and raises his hands to his head.
        Then, he slightly turns his upper body to the right. Afterward, he
        straightens his elbows and lowers his hands to his thighs. Finally,
        he straightens his elbows completely and moves his hands back to his
        sides.",

        "First, the person bends the elbows and raises his hands above his head.
        Then, he slightly rotates his upper body to the right. Subsequently, he
        straightens the elbows and lowers his hands to rest on his thighs. Fi-
        nally, he fully extends his elbows and returns his hands to their posi-
        tions at his sides.",

        "The person begins by bending the elbows and raising the hands to-
        ward the head. Subsequently, he slightly twists his upper body to the
        right. Afterward, he extends the elbows and lowers the hands toward
        the thighs, then fully straightening the elbows and moving the hands
        back to the sides."
    ],
}
```

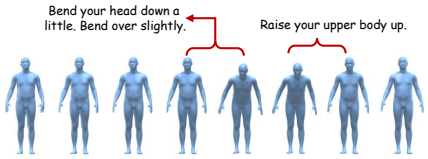
## 9. More Dataset Examples

We display more examples of body part movement descriptions for motion snippet (*i.e.*, BPMSD) and for whole motion sequence (*i.e.*, BPMP) of our FineMotion dataset in Fig. 9 and 10.



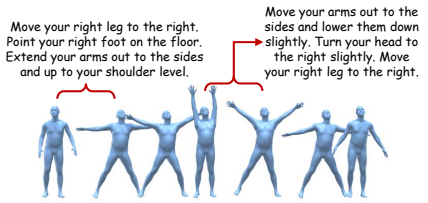

Initially, the individual advances the right leg while moving the right arm to the right and bending the elbow, followed by a slight forward movement of the left hand. Subsequently, he moves the left leg forward and raises the right hand to shoulder level, bending the elbow. The left hand then retracts slightly. Afterward, the right knee bends and moves forward, accompanied by the movement of the right arm to the left and the left hand forward. Finally, he lowers the right arm to the right side and advances the right leg, pointing the foot on the floor while retracting the left hand behind the back.

Initially, the individual moves his right leg forward and simultaneously extends his right arm to the right, bending his elbow, while moving his left hand forward slightly. Subsequently, he advances his left leg forward, lifting his right hand to shoulder level and bending his right elbow, all while moving his left hand backward slightly. Afterward, he bends his right knee and moves it forward, transitioning his right arm to the left, and moving his left hand forward. Finally, he lowers his right arm to his right side, extends his right leg forward and points his right foot on the floor, while moving his left hand backward behind his back.




The individual begins by bending his head and body slightly forward. Subsequently, he bends his body further forward. Then, he raises his upper body and head slightly. Afterward, he continues raising his upper body, maintaining an upright posture. Finally, he raises his upper body further, ensuring his body remains straight.

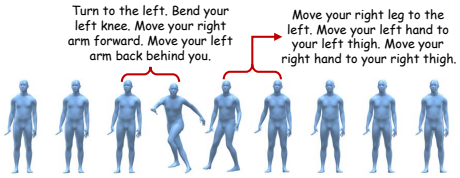
The person initially bends his head downward slightly, coupled with a slight forward bend. Subsequently, he intensifies the forward bend before gradually raising his upper body and head. Afterward, he continues raising his upper body further, maintaining an erect posture.




The individual moves his right leg to the right, pointing his foot on the floor while simultaneously extending his arms out to the sides and up to shoulder level. He continues raising his arms higher to head level and bends his knees slightly. Afterward, he moves his right leg to the left, closer to his left leg, and straightens his knees. He raises his arms up higher, moving them closer to his head. Subsequently, he moves his arms out to the sides and lowers them down slightly, turning his head to the right. He then moves his right leg to the right slightly more, points his right foot on the floor, and lowers his arms down more to chest level. He moves his right leg closer to his left leg and lowers his arms down more. Finally, he moves his arms closer to his sides, moves his right leg to the left slightly more, and points his right foot on the floor.

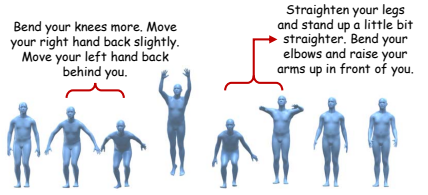
The individual moves his right leg to the right, pointing his foot towards the ground, while extending his arms laterally at shoulder height. He then raises his arms slightly higher to his head level, bending his knees concurrently. Subsequently, he moves his right leg towards the left, closer to the left leg, straightening his knees while raising his arms higher and bringing them closer to his head. Following this, he spreads his arms out to the sides, lowering them slightly, and turns his head gently to the right. He proceeds to move his right leg further to the right and points his right foot on the floor, while lowering his arms towards his chest level. Next, he brings his right leg closer to his left leg and continues lowering his arms. Lastly he moves his arms towards his sides and slightly shifts his right leg to the left, pointing his right foot towards the ground.

Figure 9. More examples of human-annotated body part movement snippet descriptions (*left*) and paragraphs (*right*). The colored text in paragraphs links to corresponding snippet descriptions.



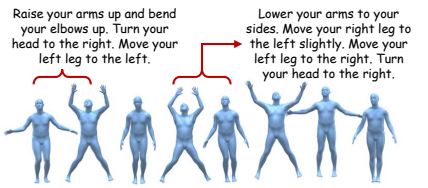

Initially, the individual turns to the left, flexing the left knee. Simultaneously, he moves his right arm forward and left arm behind him. Then, he lowers his right arm to the right thigh. Concurrently, he extends his left hand forward and shifts his right leg to the right. He turns his head to the right, followed by moving his right leg to the left, bringing his left hand to his left thigh, and positioning his right hand on the right thigh. Lastly, he moves both hands to their respective thighs.

Turning to the left, the individual bends his left knee and moves his right arm forward while simultaneously extending his left arm behind him. Subsequently, he lowers his right arm to rest on his right thigh and moves his left hand forward. As he moves his right leg to the right, he turns his head in the same direction. Continuing to move his right leg to the left, he positions his left hand on his left thigh and his right hand on his right thigh. Finally, he repeats the motion of placing his right hand on his right thigh and his left hand on his left thigh.




The individual initiates the sequence by moving the left hand to the left and the right hand to the right, accompanied by a turn of the head to the right. He then bends his knees further while retracting the left hand slightly and moving the right hand behind him. Afterward, he straightens his legs and raises both arms in front, directing his gaze upward and to the left. Consequently, he bends his knees again and lowers his arms to his sides, keeping his head turned to the right. The individual proceeds to straighten his legs and stand upright, followed by bending his elbows and raising his arms in front of him. He then lowers his arms to his sides and turns his head to the right. Subsequently, he extends his right hand forward and places his left hand on his left hip. Leaning his upper body to the left, he shifts his left hand to his left hip and his right hand to his right thigh.

Initially, the individual moves his left hand to the left and his right hand to the right while turning his head to the right. Next, bends his knees and slightly retracts his left hand before moving his right hand back behind him. He then straightens his legs, raises his arms in front, looks up and to his left. Subsequently, he bends his knees, lowers his arms, and turns his head to the right before straightening his legs, standing straighter, bending his elbows, and raising his arms again. The person lowers his arms, turns his head to the right, moves his right hand forward, and positions his left hand on his left hip. He leans his upper body to the left while keeping his left hand on his left hip and moving his right hand to his right thigh.




The individual begins by raising his arms and bending his elbows, then turns his head to the right and moves his left leg to the left. Subsequently, he lowers his arms to his sides, specifically the right arm to its right side and slightly shifts his left leg to the right. Next, he re-raises his arms and bends his elbows, turns his head to the right, and moves his left leg to the left with a slight bend in the knee. Again, he lowers his arms to his sides and makes minor adjustments by moving both legs and turning his head to the right. Afterward, he extends both arms out to his sides, turns his head to the right, extends his left leg to the left with a pointed foot, and rotates his right foot to the right. Finally, he lowers his arms to his sides, turns his head to the right and moves his left leg to the right, before lowering his arms and turning his head to the left and moving his right foot to the right.

Initially, the person raises his arms and bends his elbows upward, while simultaneously turning his head to the right and moving his left leg to the left. He then lowers his arms to his sides and continues lowering his right arm to the right side, concurrently shifting his left leg slightly to the right. Subsequently, he raises his arms again, bends his elbows, and turns his head to the right, while moving his left leg to the left and bending his left knee. Afterward, he lowers his arms to his sides, moves his right leg slightly to the left, and shifts his left leg to the right, all the while maintaining his head turned to the right. He then raises his arms to the sides and extends his left leg to the left, pointing his left foot on the floor, while turning his right foot to the right. Finally, he lowers his arms, turns his head to the right, moves his left leg to the right, and turns his head to the left, shifting his right foot to the right as well.

Figure 10. More examples of automatically generated body part movement snippet descriptions (*left*) and paragraphs (*right*). The colored text in paragraphs links to corresponding snippet descriptions.

## 10. Baseline Model Details

This section outlines the network architecture and the implementation of three variants of motion generation methods, including MDM [25], T2M-GPT [30], and MoMask [8] on our dataset, and denoted them as (T&DT)-MDM, (T&DT)2M-GPT, and (T&DT)-MoMask, respectively.

- **(T&DT)-MDM** builds from MDM [25]. It employs a classifier-free, diffusion-based approach for human motion generation using a transformer-based architecture. Unlike standard diffusion models, it directly predicts the sample at each diffusion step. Specifically, the transformer-encoder predicts the final clean motion based on a condition (*i.e.*, a CLIP-based textual embedding), a noising timestep, and random noise. To accommodate detailed textual descriptions, which often contain over ten times the number of tokens compared to coarse captions, we replace the CLIP [21] text encoder with the T5-Base [22] encoder, which uses relative attention for flexible input lengths. We then perform mean pooling along the sequence length dimension of the T5-Base encoder output to obtain a single text embedding for each text. Now, the model's condition turns out to be the concatenated text embeddings of both the coarse caption and the detailed description.
- **(T&DT)2M-GPT** is derived from T2M-GPT [30] and comprises a Motion VQ-VAE and a GPT model. Motion VQ-VAE learns a mapping between raw motion sequences and discrete token sequences, while the GPT model generates motion tokens conditioned on text embeddings. Likewise, we modified the condition of the GPT model into the concatenated T5 text embeddings of the coarse caption and the detailed text.
- **(T&DT)-MoMask** is based on MoMask [8], featuring a Motion Residual VQ-VAE, a Masked Transformer, and a Residual Transformer. Concretely, the Residual VQ-VAE uses a hierarchical quantization scheme to discretize motions into multiple layers of motion tokens. The Masked Transformer predicts masked motion tokens from the text input, while the Residual Transformer progressively predicts next-layer tokens based on the results from the current layer. The textual embedding is modified similarly to the previous two networks.

All three baseline models are adapted to include our long, detailed body part movement descriptions for the motion sequences. Here, we hold (T&DT)2M-GPT as the example to elaborate on the differences from the original T2M-GPT model. The modifications applied to the other two baseline models, (T&DT)-MDM and (T&DT)-MoMask, follow a similar approach.

(T&DT)2M-GPT mainly contains two parts: Motion VQ-VAE for motion discretization and GPT for generating motion tokens from the coarse caption and detailed text.

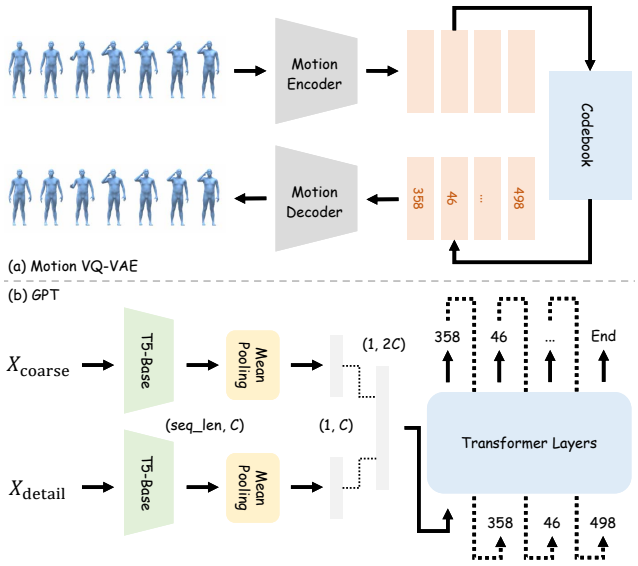



Figure 11. **Overview of the baseline network, (T&DT)2M-GPT.** It generates motions that strictly follow the fine-grained description $X_{\text{detail}}$ and the coarse-grained caption $X_{\text{coarse}}$. It consists of a motion VQ-VAE for discretizing the motion into tokens and a GPT for generating motion tokens.

**Motion VQ-VAE.** We follow [30] to represent motions in discrete tokens, and vice versa. Specifically, it contains an encoder $E$, a decoder $D$, and a learnable codebook $B = \{b_k\}_{k=1}^{K}$, where $K$ is the size of the codebook. Given a $T$-frame motion sequence $M = [m_1, m_2, \ldots, m_T]$ with $m_t \in \mathbb{R}^d$, the encoder $E$ maps it into a sequence of latent features $Z = E(M)$ with $Z = [z_1, z_2, \ldots, z_{\lfloor T/l \rfloor}]$ and $z_i \in \mathbb{R}^{d_c}$, where $l$ represents the temporal downsampling rate of the encoder $E$. Then, these latent features are transformed into a sequence of motion codes $C = [c_1, c_2, \ldots, c_{\lfloor T/l \rfloor}]$, where $c_i$ is the index of the most similar element to $z_i$ in $B$. With a sequence of motion codes $C$, we first project $C$ back to their corresponding codebook elements $\widetilde{Z} = [\widetilde{z}_1, \widetilde{z}_2, \ldots, \widetilde{z}_{\lfloor T/l \rfloor}]$ with $\widetilde{z}_i = b_{c_i}$. Then, the decoder $D$ reconstructs $\widetilde{Z}$ into a motion sequence $\widetilde{M} = D(\widetilde{Z}) = [\widetilde{m}_1, \widetilde{m}_2, \ldots, \widetilde{m}_T]$. The motion VQ-VAE is optimized by the standard optimization goal [27] that requires the decoded motion $\widetilde{M}$ to be as similar as the input motion $M$. The exponential moving average (EMA) and codebook reset (Code Reset) are employed to stabilize the training process. With a learned motion VQ-VAE, a motion sequence can be easily mapped into discrete motion tokens by the encoder $E$ and the codebook $B$. On the other hand, the output of our (T&DT)2M-GPT model, *i.e.*, motion tokens, can be recovered into motion sequences by the decoder $D$ and the codebook $B$.

**GPT for generating motion tokens.** First, we extract the text embeddings of the coarse caption $t_{\text{coarse}}$ and the edited detailed motion script $\hat{t}_{\text{detail}}$. Since the number of tokens of our detailed human body part descriptions is usually more than ten times that of the coarse captions, we use the frozen encoder from T5-Base [22] to extract the textual embeddings, considering that its relative attention mechanism allows input with any sequence length. We then perform a mean pooling operation in the *seq_len* dimension of the output from the T5-Base encoder to obtain a single text embedding for each text.

$$t_{\text{coarse}} = \text{Mean}(\text{T5Encoder}(X_{\text{coarse}})) \in \mathbb{R}^{768}, \quad (1)$$

$$\hat{t}_{\text{detail}} = \text{Mean}(\text{T5Encoder}(\hat{X}_{\text{detail}})) \in \mathbb{R}^{768}. \quad (2)$$

Next, the two text embeddings are utilized as the conditions of the GPT model to autoregressively generate motion tokens. The GPT model is composed of a stack of transformer layers. Besides, casual self-attention is applied to ensure the calculation of the current tokens does not consider the information of the future motion tokens. Since this fine-grained motion generation task can be considered as the next motion token prediction task, which is based on the given coarse textual embedding, the motion script textual embedding, and previous motion tokens, the GPT model is optimized by the cross-entropy loss between the predicted motion tokens and ground-truth ones.

$$L = -\sum_{i=1}^{\lfloor T/l \rfloor} log(P(c_i \mid t_{\text{coarse}}, \hat{t}_{\text{detail}}, c_{<i}, \theta_{\text{GPT}})). \quad (3)$$

After sufficient training, the GPT model can generate appropriate motion tokens that can be further decoded into motions by the decoder in Motion VQ-VAE.

## 11. More Implementation Details

The architecture and training hyperparameters of our baseline models strictly follow those in the original paper [8, 25, 30]. Notably, since we replace the text encoder with that of T5, the dimension of output text embedding turns to 768 rather than that of the CLIP text encoder, 512. Therefore, the input of the fully connected layer that projects the CLIP text embedding to the input of the GPT also needs to be changed from 512 to 768. The code is based on PyTorch. The experiments were conducted on the A100-80G GPU, but only about 16G GPU memory was used. Due to replacing the text encoder with a larger model [22] and using it to process longer textual descriptions, the training time for (T&DT)2M-GPT increases to 154 hours, compared to the 78 hours reported in [30] for T2M-GPT. However, the training time can be reduced to the original 78 hours if all text embeddings are pre-extracted and stored before the training begins.

## 12. More Discussion on Motion Generation with Fine-grained Texts Only

We did not evaluate this setting because it will lead to **ambiguity** in motion generation. Fine-grained text captures detailed body part movements and timing, while coarse text supplements global motion semantics, both crucial for precise motion generation. For instance, motions with coarse text '*a person is standing still*' and '*a person is sitting*' share the same fine-grained text (<Motionless>, *i.e.*, no body part movements). The model cannot distinguish such cases without coarse text, degrading motion generation performance. Given the issue above, we do not train our models using (fine-grained text, motion) pairs. Evaluating such a setting without proper training would lead to unfair or unreliable results.

## 13. More Discussion on Table 2

One may notice that when (T&DT)2M-GPT—*i.e.*, Rows (2)-(5) in Table 2—generates motions using only coarse descriptions (Test Set: T2M), it shows a slight performance drop, compared to our implementation of T2M-GPT trained solely on the T2M task, Row (1). The slight drop in T2M-GPT variants likely stems from their high sensitivity to the shared training budget, as multi-task training with (T&DT) halves the T2M updates compared to the baseline. Additional evaluations on MDM and MoMask variants show that including (T&DT)2M during training actually improves motion generation when only coarse text is available, as shown below.

| Train Task | | Test Task | MDM | | T2M-GPT | | MoMask | |
|---|---|---|---|---|---|---|---|---|
| T2M | (T&DT)2M | T2M | R-Top3 ↑ | FID ↓ | R-Top3 ↑ | FID ↓ | R-Top3 ↑ | FID ↓ |
| ✓ | - | ✓ | $0.606^{\pm.008}$ | $3.137^{\pm.183}$ | $0.781^{\pm.003}$ | $0.123^{\pm.005}$ | $0.753^{\pm.002}$ | $0.249^{\pm.012}$ |
| ✓ | (our BPMSD) | ✓ | $0.746^{\pm.007}$ | $0.760^{\pm.064}$ | $0.781^{\pm.002}$ | $0.154^{\pm.007}$ | $0.827^{\pm.002}$ | $0.120^{\pm.004}$ |
| ✓ | (our BPMP) | ✓ | $0.759^{\pm.006}$ | $0.436^{\pm.043}$ | $0.781^{\pm.002}$ | $0.155^{\pm.006}$ | $0.818^{\pm.002}$ | $0.130^{\pm.005}$ |

Table 4. Generation performance of all our variants on the T2M test set, *i.e.*, motion generation conditioned on coarse descriptions only.

## 14. Ablation Study on Baseline Model Design

Here, we conduct an ablation study on different strategies for encoding coarse and detailed texts. Specifically, we denote the strategy of connecting the coarse text (T) and detailed text (DT) into a single text and feeding it into the text encoder as 'TDT'. Meanwhile, 'T&DT' refers to encoding T and DT separately and then concatenating their resulting embeddings. Results below show that the 'TDT' strategy leads to poorer performance, likely because the model is overwhelmed by the dense information and struggles to capture the global motion semantics. These findings highlight that our baseline designs are carefully considered, rather than naïve implementations.

| Method | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| TDT-MoMask (BPMSD) | $0.212^{\pm.002}$ | $0.341^{\pm.002}$ | $0.434^{\pm.002}$ | $8.328^{\pm.056}$ | $5.877^{\pm.009}$ | $8.899^{\pm.069}$ |
| (T&DT)-MoMask (BPMSD) | $0.519^{\pm.002}$ | $0.715^{\pm.002}$ | $0.811^{\pm.001}$ | $0.088^{\pm.003}$ | $2.946^{\pm.005}$ | $9.702^{\pm.075}$ |
| TDT-MoMask (BPMP) | $0.358^{\pm.003}$ | $0.528^{\pm.002}$ | $0.628^{\pm.002}$ | $0.285^{\pm.006}$ | $4.145^{\pm.008}$ | $9.626^{\pm.093}$ |
| (T&DT)-MoMask (BPMP) | $0.520^{\pm.003}$ | $0.717^{\pm.002}$ | $0.813^{\pm.002}$ | $0.055^{\pm.002}$ | $2.935^{\pm.009}$ | $9.679^{\pm.085}$ |

Table 5. Ablation study on different strategies for encoding coarse and detailed texts.

## 15. Metrics and Results for Temporal Alignment

Currently, there is no metric that directly evaluates the precision of temporal alignment between detailed texts and generated motion sequences. Given that our detailed texts are strictly aligned with ground-truth motions over time, we reframe this evaluation as measuring the alignment between short clips of generated motions and corresponding ground-truth clips. High similarity between these clips—even at fine temporal granularity—implies accurate alignment with the detailed texts.

To this end, we introduce $\text{FID}_c$, which computes the similarity between generated and ground-truth motions using overlapping 40-frame clips (the minimum evaluation length), with a stride of 10—matching the minimal temporal interval of our detailed texts. The table below reports $\text{FID}_c$ scores across all clips. As shown, our variants (last two rows) achieve significantly lower $\text{FID}_c$ scores, demonstrating that our generated motions are better temporally aligned with the detailed texts, compared to motions generated by models trained solely on coarse descriptions.

| | MDM | T2M-GPT | MoMask |
|---|---|---|---|
| T2M | $3.012^{\pm.206}$ | $1.423^{\pm.040}$ | $0.293^{\pm.011}$ |
| (T&DT)2M (BPMSD) | $1.382^{\pm.125}$ | $0.398^{\pm.011}$ | $0.165^{\pm.004}$ |
| (T&DT)2M (BPMP) | $0.426^{\pm.046}$ | $0.624^{\pm.015}$ | $0.108^{\pm.003}$ |

Table 6. Comparison of temporal alignment, measured by $\text{FID}_c$, between baseline text-to-motion models and our fine-grained variants.

## 16. Limitations and Future Work

Since we use temporally augmented data to train the text-to-motion models, editing motions along the temporal dimension becomes more straightforward and accurate compared to spatial editing. Consequently, future work will focus on developing effective methods for spatial human motion editing.

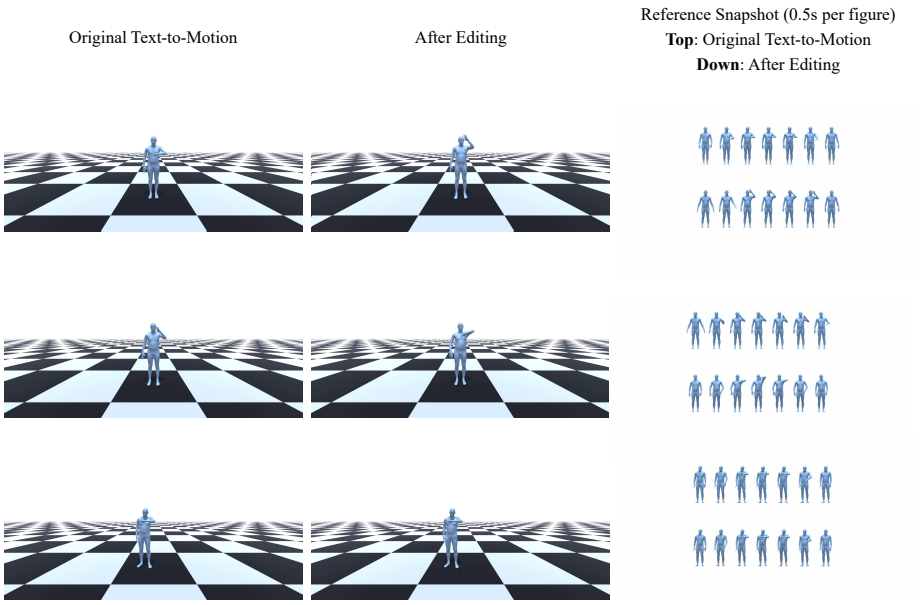
Additionally, obtaining the detailed body part textual descriptions still requires multiple steps. Thus, training an end-to-end model that can directly infer these descriptions from human motion sequences presents a promising research direction.

Moreover, the capabilities of large language models (LLMs) could be leveraged to unify text-to-motion and motion-to-text tasks through textual descriptions of varying granularity, potentially enhancing the effectiveness of both tasks.

## 17. User Study

**Case 1**: **Add** the body part movements **Spatially**.

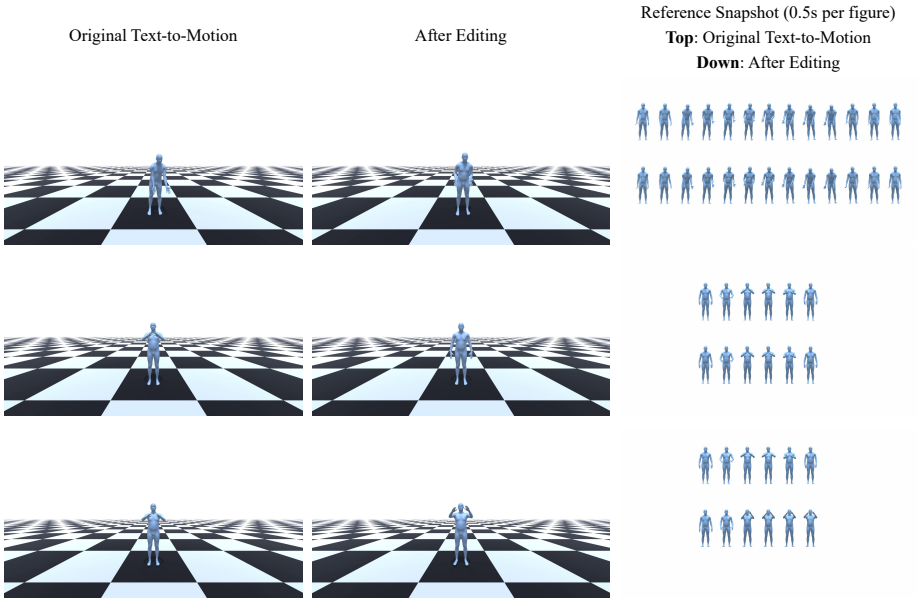
original text: a person lifts their left wrist towards their face as if to look at a watch
editing requirement: Lift your left hand to the head.




Answer: Row 1: Ours Row 2: T2M-GPT Row 3: FLAME

**Case 2**: **Delete** the body part movements **Spatially**.

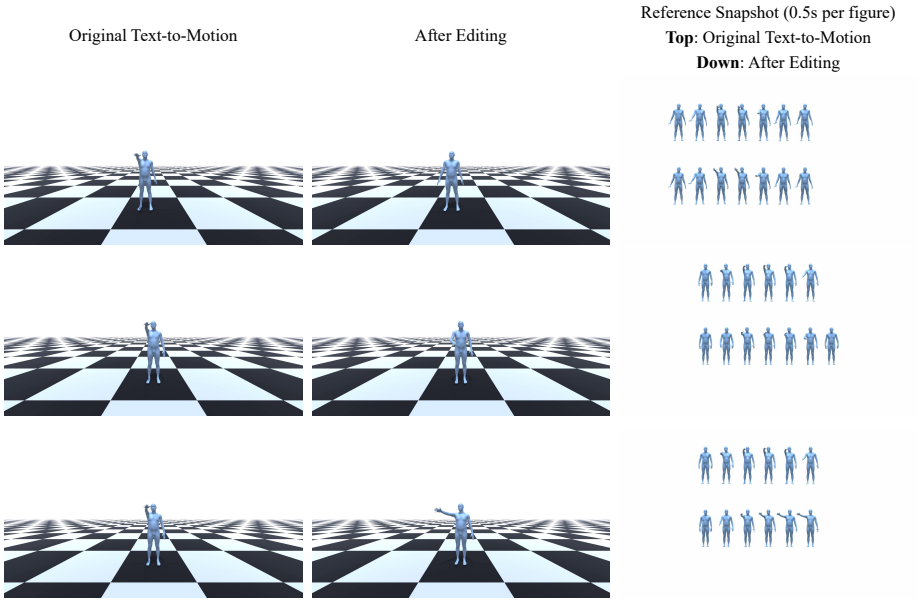
original text: a man bends his arms to touch an object in front of him.
editing requirement: Do not put your hands down to your sides.




Answer: Row 1: FLAME Row 2: T2M-GPT Row 3: Ours

**Case 3**: **Modify** the body part movements **Spatially**.

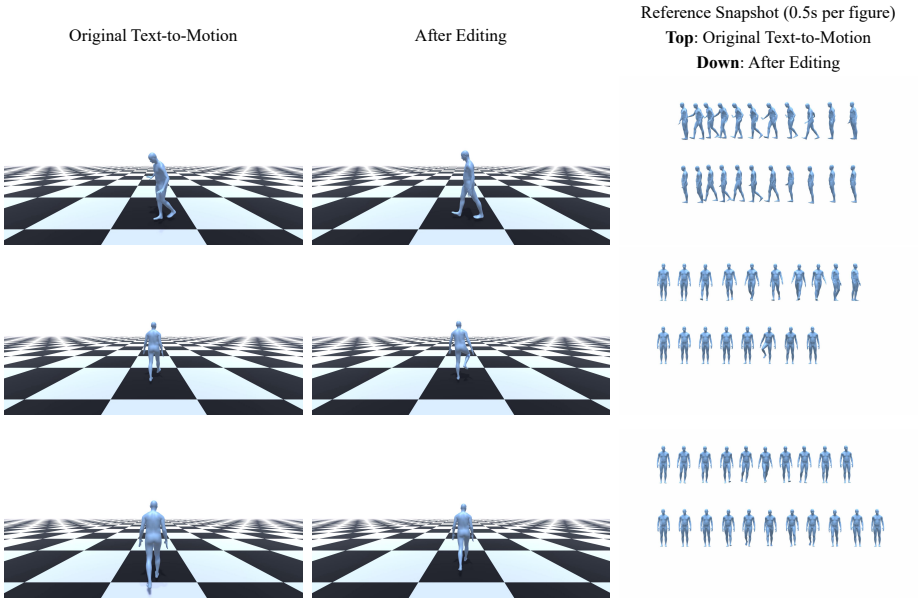
original text: a person stretches right hand forward
editing requirement: Do not put down your right hand.




Answer: Row 1: FLAME Row 2: T2M-GPT Row 3: Ours

**Case 4**: **Extend** at the **start** of the human motion (**Temporally**).

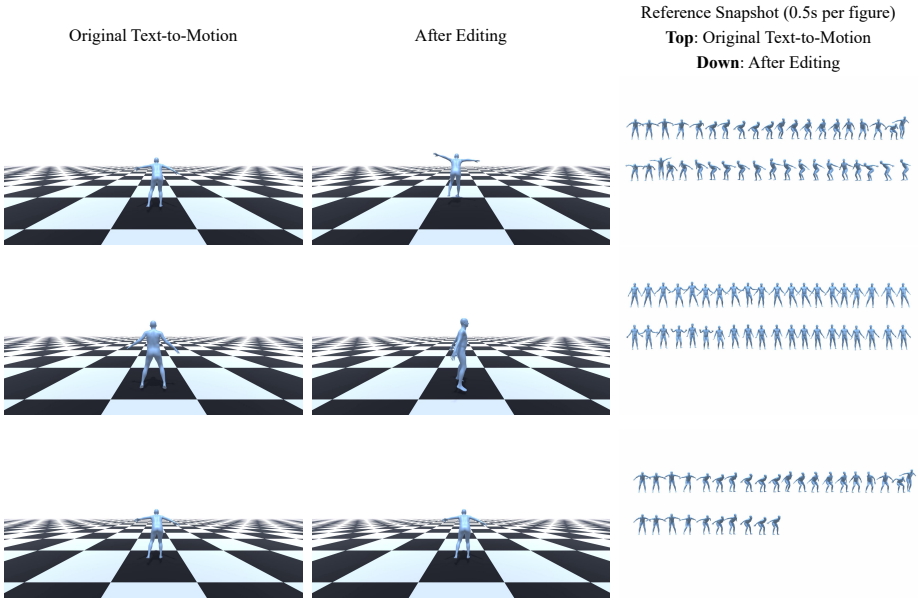
original text: a person walks forward while making small adjustments left and right
editing requirement: Stand for one second before start walking.




Answer: Row 1: FLAME Row 2: T2M-GPT Row 3: Ours

**Case 5**: **Delete** at the **end** of the human motion (**Temporally**).

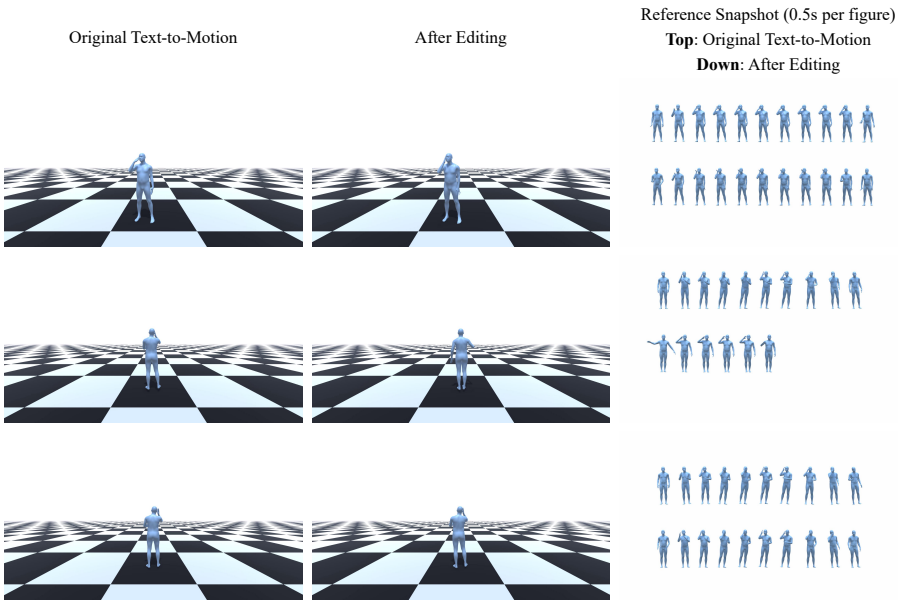
original text: a person hops with both feet in a half circle while both arms are positioned backwards.
editing requirement: Delete the motion in the last 4 seconds.




Answer: Row 1: T2M-GPT Row 2: FLAME Row 3: Ours

**Case 6**: **Delete** in the **middle** of the human motion (**Temporally**).

original text: a person leaned something near to face with right hand
editing requirement: Delete the motion from 1.0-2.5s.

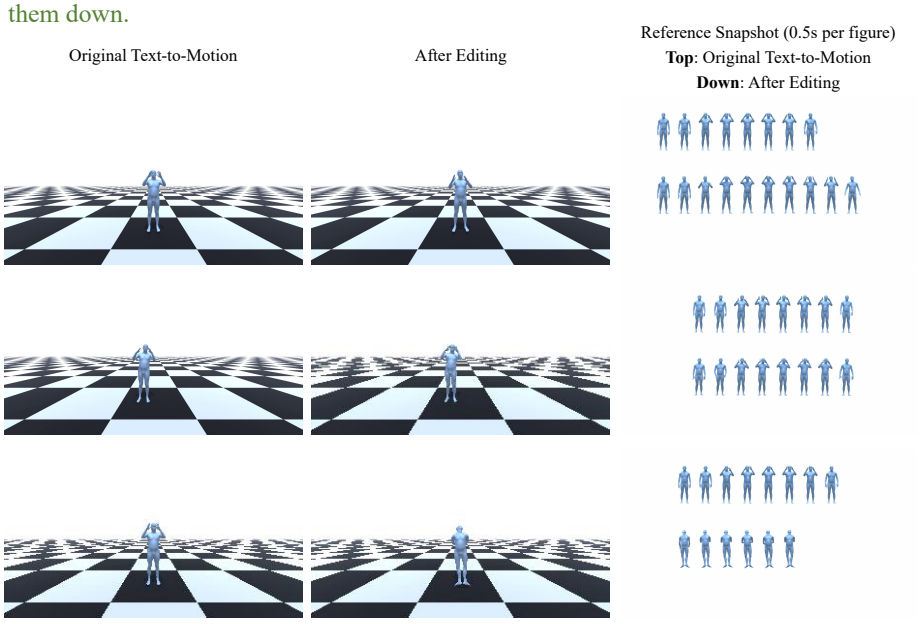



Answer: Row 1: FLAME Row 2: Ours Row 3: T2M-GPT

**Case 7**: **Insert** in the middle of the human motion (**Temporally**).

original text: a person lifts both hands toward face and then lowers them to their sides.

editing requirement: After lifting your hands, stay for one more second, and then lower them down.

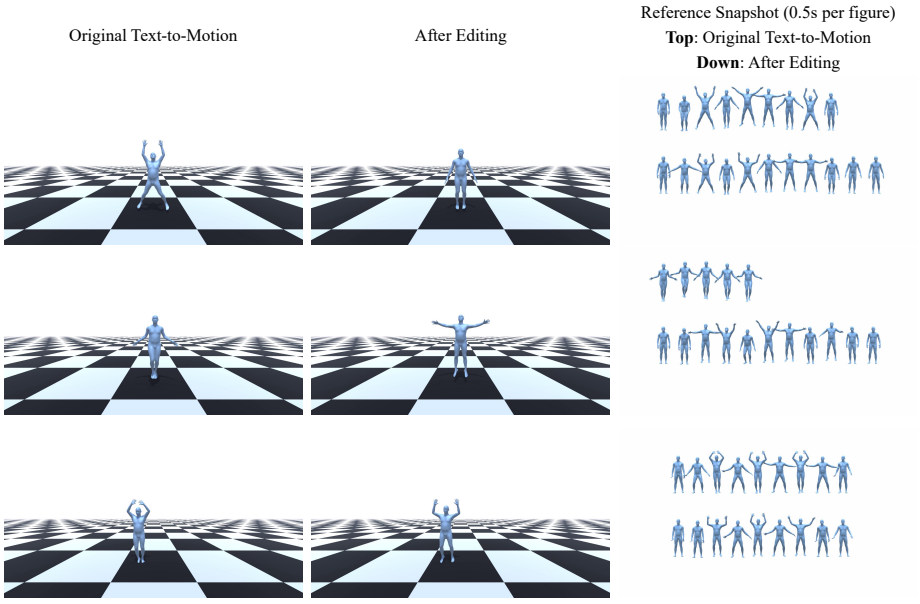



Answer: Row 1: Ours Row 2: FLAME Row 3: T2M-GPT

**Case 8**: **Extend** at the **end** of the human motion (**Temporally**).

original text: a person does jumping jacks.

editing requirement: Stand for one more second after jumping.

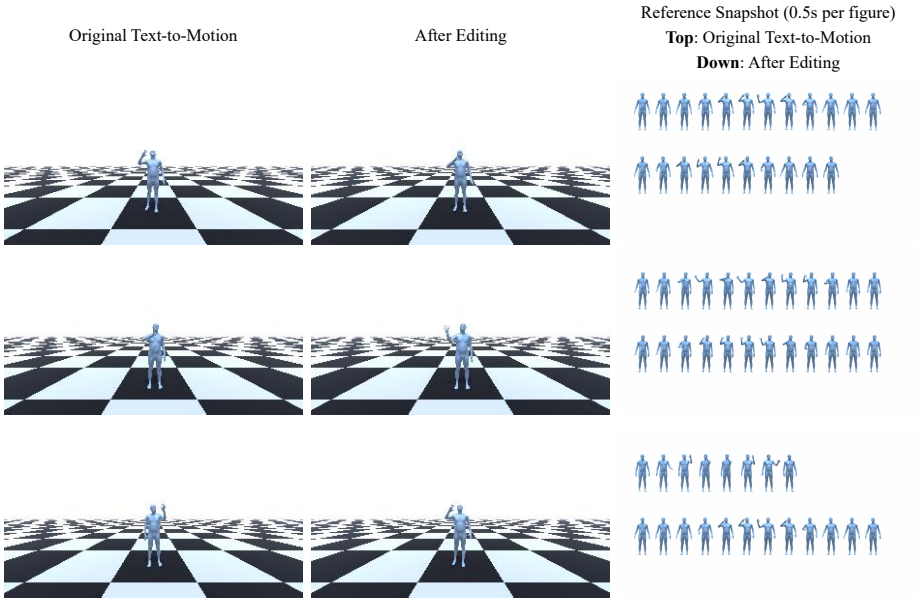



Answer: Row 1: Ours Row 2: T2M-GPT Row 3: FLAME

**Case 9**: **Delete** at the **start** of the human motion (**Temporally**).

original text: a man waves his right hand.

editing requirement: Delete the standing still segment at the start of the motion.




Answer: Row 1: Ours Row 2: FLAME Row 3: T2M-GPT