FineXtrol: Controllable Motion Generation via Fine-Grained Text

Keming Shen[1,2], **Bizhu Wu**[1,2,3], **Junliang Chen**[4], **Xiaoqin Wang**[1,2], **Linlin Shen**[1,2,5*]
[1]School of Computer Science and Software Engineering, Shenzhen University
[2]Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University
[3]School of Computer Science, University of Nottingham Ningbo China, Ningbo, China
[4]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University
[5]Computer Vision Institute, School of Artificial Intelligence, Shenzhen University
2400101005@mails.szu.edu.cn, llshen@szu.edu.cn

Abstract

Recent works have sought to enhance the controllability and precision of text-driven motion generation. Some approaches leverage large language models (LLMs) to produce more detailed texts, while others incorporate global 3D coordinate sequences as additional control signals. However, the former often introduces misaligned details and lacks explicit temporal cues, and the latter incurs significant computational cost when converting coordinates to standard motion representations. To address these issues, we propose FineXtrol, a novel control framework for efficient motion generation guided by temporally-aware, precise, user-friendly, and fine-grained textual control signals that describe specific body part movements over time. In support of this framework, we design a hierarchical contrastive learning module that encourages the text encoder to produce more discriminative embeddings for our novel control signals, thereby improving motion controllability. Quantitative results show that FineXtrol achieves strong performance in controllable motion generation, while qualitative analysis demonstrates its flexibility in directing specific body part movements.

1 Introduction

Human motion generation (Guo et al. 2020; Harvey et al. 2020; Song et al. 2021; Chen et al. 2023; Jiang et al. 2023) has become an essential task for various applications such as animation and digital humans. Recent advancements (Guo et al. 2022b,a; Zhang et al. 2024; Tevet et al. 2023; Guo et al. 2024; Tan et al. 2024; Wang et al. 2025a) have shown impressive capability in synthesizing diverse and realistic human motions from given texts, as illustrated in Fig. 1(A). As the quality of generated motions improves, there is a growing demand for generating motions that are not only realistic but also **precise** and **controllable**. Existing works to address these demands fall into two main categories.

On the one hand, some methods (Kalakonda, Maheshwari, and Sarvadevabhatla 2023; Huang et al. 2024; Wang et al. 2025c) leverage detailed textual descriptions to generate precise motions, as shown in Fig. 1(B). Specifically, large language models (LLMs) are used to expand the coarse-grained textual annotations (*e.g.*, "A man stands still and he is playing the violin.") into descriptions detailing the movements of individual body parts. These expanded fine-grained descriptions are then connected with the original coarse texts to generate motions. However, this approach suffers from several limitations. First, **the expanded descriptions are often not strictly aligned with the ground-truth motions**, as noted in (Wu et al. 2025), and may contain inaccurate information due to LLM bias. Moreover, the expanded descriptions detail all body part movements of the whole motion sequence, **lacking explicit temporal cues**, such as *when to raise a hand*, making it difficult to achieve precise control over motion within specific time intervals. Besides, this **simple connection paradigm provides limited flexibility for temporally localized control**, as shown in Tab. 4.

On the other hand, methods like (Xie et al. 2024; Wang et al. 2024) introduce spatial control signals to improve controllability, as illustrated in Fig. 1(C). Inpainting-based methods (Shafir et al. 2024; Tevet et al. 2023; Karunratanakul et al. 2023) integrated spatial constraints into the generation process, but struggle to control joints beyond the pelvis due to the use of relative pose representations. Recent methods (Xie et al. 2024; Wang et al. 2024) addressed the above limitations by converting the generated motion to global coordinates, allowing direct comparison with input control signals and using gradient-based error refinement. However, this coordinate transformation **incurs significant computational cost**, and specifying realistic global coordinate sequences is challenging and unintuitive for users, **limiting applicability and scalability**.

To combine the strengths of both categories while addressing their limitations, we propose a novel controllable motion generation framework: FineXtrol—Fine-grained teXt controllable motion generation, as displayed in Fig.1(D). Instead of relying on spatial coordinate sequences, FineXtrol uses fine-grained textual descriptions of individual body part movements (*e.g.*, "*Move your left hand to your left thigh.*") as the control signals. This eliminates the need for coordinate conversion, improving efficiency, and offers better scalability through the more user-friendly form of control. Moreover, our fine-grained textual control signals are sourced from FineMotion (Wu et al. 2025), whose descriptions are explicitly aligned with ground-truth motions and encode explicit temporal information. Besides, rather

*Corresponding author.

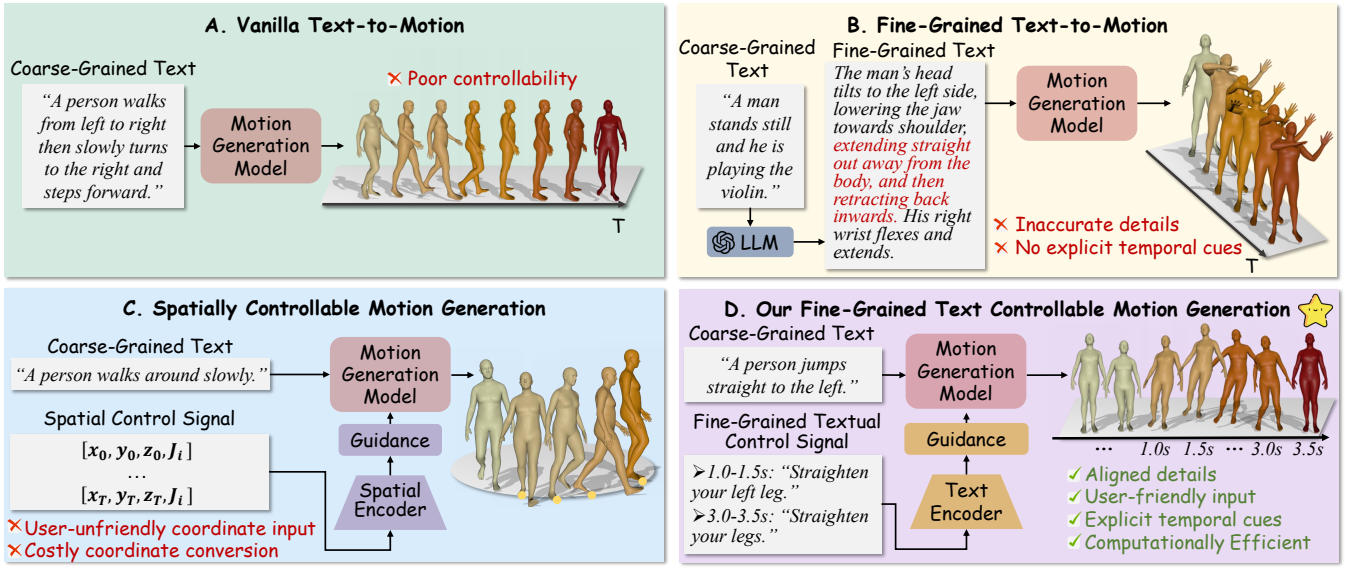

Figure 1: Illustrations of (A) *Vanilla text-to-motion* methods struggle to control specific body part movements. (B) *Fine-grained text-to-motion* approaches using LLMs' expanded descriptions for fine details, but often misalign with ground-truth motions and lack explicit temporal cues. (C) *Spatially controllable motion generation* methods rely on global 3D coordinate sequences as extra control signals, which are difficult to be provided beyond existing datasets and incur high computational costs from pose conversion. (D) Our *FineXtrol* introduces accurate and temporally explicit fine-grained textual control signals for specific body parts, enabling user-friendly and efficient controllable motion generation.

than directly connecting fine-grained textual control signals with coarse-grained texts as a single input, we incorporate the control signals as residual guidance to modulate the motion features conditioned on coarse-grained texts. These designs lead to the generation of realistic and coherent motions that adhere closely to the specified fine-grained constraints.

During implementation, we observed that widely used text encoders such as CLIP (Radford et al. 2021) and T5 (Raffel et al. 2020) struggle to produce discriminative embeddings for our fine-grained textual control signals, as illustrated in Fig. 7. We attribute this limitation to their pre-training on datasets that emphasize coarse semantics. As a result, these encoders often overlook subtle cues essential for capturing detailed motion semantics. To address this issue, we analyze the characteristics of our fine-grained textual control signals, and introduce a hierarchical contrastive learning module. This tailored module enhances the text encoder's ability to extract discriminative embeddings for our fine-grained textual control signals, thereby enabling more precise control over human motion generation.

Both quantitative results on HumanML3D (Guo et al. 2022a) and visualizations under various settings demonstrate that our framework, FineXtrol, delivers strong controllable motion generation performance and notably surpasses the previous state-of-the-art in controlling multiple body parts within designated temporal intervals. Moreover, compared with prior diffusion-based controllable motion generation methods, FineXtrol uses fewer trainable parameters and requires less inference time, highlighting the efficiency of our framework. To summarize, our contributions are as follows: **First**, to the best of our knowledge, we propose FineXtrol, a novel and user-friendly controllable motion generation framework that enables control over body part movements within specific temporal intervals. **Secondly**, we design a hierarchical contrastive learning module to enhance the text encoder's ability to capture discriminative embeddings for our fine-grained textual control signals, thereby further improving FineXtrol's capacity to generate motions that accurately follow the specified instructions. **Thirdly**, extensive experiments demonstrate that our FineXtrol excels in precise multi-body-part motion control, and user accessibility, highlighting its potential for real-world applications.

2 Related Work

Text-driven Human Motion Generation. Text-to-motion generation plays a crucial role in various applications. Existing approaches can be broadly categorized into two major generative paradigms (Xie et al. 2024): auto-regressive models (Guo et al. 2022a, 2024; Starke, Mason, and Komura 2022; Rempe et al. 2023; Huang et al. 2024; Juravsky et al. 2022) and diffusion-based models (Tevet et al. 2023; Yuan et al. 2023; Shafir et al. 2024; Karunratanakul et al. 2023; Chen et al. 2023). Auto-regressive approaches generate motion frame-by-frame, while diffusion-based ones progressively denoise an entire motion sequence over multiple iterations. Most of them rely on coarse-grained texts, which often fail to capture the subtle details of human actions. Recent studies (Kalakonda, Maheshwari, and Sarvadevabhatla 2023; Huang et al. 2024; Wang et al. 2025c) have incorporated fine-grained textual descriptions to enhance precise motion generation. However, their LLM-generated detailed texts are often

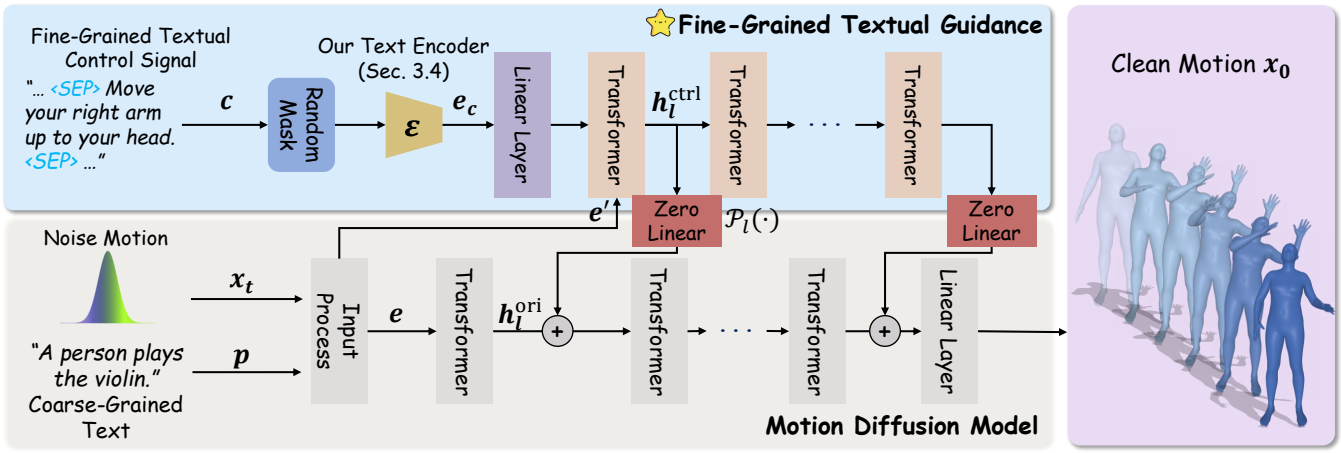

Figure 2: Overview of FineXtrol. Our framework takes the coarse-grained text $\boldsymbol{p}$, the fine-grained textual control signal $\boldsymbol{c}$, and a noise motion sequence $\boldsymbol{x_t}$ as input, and predicts the clean motion sequence $\boldsymbol{x_0}$. The lower branch resumes from MDM to maintain stable motion generation capabilities from $\boldsymbol{p}$. The upper branch is a trainable copy of MDM, modulated by $\boldsymbol{c}$ through conditional feature adaptation. The zero-initialized linear layers connect between branches. The framework ensures that the generated motion adheres not only to the coarse-grained text but also to the control signal.

not strictly aligned with ground-truth motions and lack explicit temporal cues, leading to weak correspondence between text and motion segments. This limits control over individual body parts and time intervals. In contrast, our FineXtrol enables precise control over both.

Controllable Motion Diffusion Models. Controllability in diffusion models has been widely studied, especially in image synthesis (Rombach et al. 2022; Ho and Salimans 2022; Zhang, Rao, and Agrawala 2023; Wu et al. 2024), and is now also gaining attention in motion generation area. Inpainting-based methods like PriorMDM (Shafir et al. 2024) condition on observed motion segments to predict missing ones while maintaining global coherence. ControlNet-based methods enhance controllability by injecting conditional signals via a trainable copy of a pretrained motion generation model. For example, OmniControl (Xie et al. 2024) and InterControl (Wang et al. 2024) use global joint coordinates as control signals, enabling controllability with minimal fine-tuning. However, they require users to provide realistic coordinate sequences, which is impractical, and incur expensive conversions to relative motion representations. Similarly, our method adopts ControlNet paradigm, but employs a more efficient and user-friendly control signal, *i.e.*, fine-grained textual descriptions of body part movements across temporal intervals.

Contrastive Learning for Text Embedding. Learning discriminative sentence embeddings is fundamental to natural language processing. Early methods relied on distributional hypotheses to predict surrounding sentences (Kiros et al. 2015; Logeswaran and Lee 2018; Petrovich, Black, and Varol 2021) or extended Word2Vec with N-Gram representations (Pagliardini, Gupta, and Jaggi 2018). Recent methods have adopted contrastive learning to align augmented sentence pairs (Devlin et al. 2019; Giorgi et al. 2021; Carlsson et al. 2021; Yan et al. 2019; Gao, Yao, and Chen 2021). Contrastive learning has also been effective in multi-modal settings (Wang et al. 2022; Ramesh et al. 2021; Radford et al. 2021; Wang et al. 2025b; Li et al. 2024), aligning image and text embeddings from large-scale paired data. However, most pretrained text encoders focus on coarse semantics and struggle with fine-grained distinctions. To address this, we analyze the properties of our fine-grained textual control signals, and design a contrastive module to extract discriminative embeddings, enabling more precise fine-grained text controllable motion generation.

3 Method

Fig. 2 displays the overall pipeline of our proposed framework, which leverages fine-grained textual descriptions as control signals to guide motion generation. The notations and the definition of this novel research problem are clarified in Sec. 3.1. We then briefly introduce required preliminaries in Sec. 3.2. Next, we detail two key modules in our framework: (1) generating motion sequences guided by our fine-grained textual control signals, which detail specific body part movements in specific temporal intervals (see Sec. 3.3), and (2) representing our fine-grained textual control signals in a robust and discriminant manner (see Sec. 3.4).

3.1 Problem Formulation

Given a coarse-grained text $\boldsymbol{p}$, such as "*A man kicks something with his left leg*", and a fine-grained textual control signal $\boldsymbol{c}$, such as "*Move your left leg to the right in 1.0-1.5s*", our framework $\mathcal{F}$ should generate a realistic motion sequence $\boldsymbol{x_0} \in \mathbb{R}^{T \times D}$, where T is the temporal length and $\mathcal{D}$ is the dimension of human pose representations. Mathematically, we formulate this fine-grained text controllable motion generation task as:

$$\boldsymbol{x_0} = \mathcal{F}\left(\boldsymbol{p}, \boldsymbol{c}; \Theta\right), \tag{1}$$

where Θ is the parameters of $\mathcal{F}$.

3.2 Preliminaries

Motion Diffusion Model (MDM) (Tevet et al. 2023) adapted powerful diffusion models to human motion generation by framing the task as a gradual denoising process. Given a ground-truth motion $\hat{\boldsymbol{x}}_{\mathbf{0}}$, noise is progressively added to produce a noisy version $\boldsymbol{x}_{\boldsymbol{t}}$, where t denotes the number of noise addition steps. Then, conditioned on a time step t and a coarse-grained text $\boldsymbol{p}$, MDM employed a transformer-based network ϵ_θ with parameters θ to reverse the process and directly predict the clean motion $\boldsymbol{x}_{\mathbf{0}}$. The model is trained with the objective function:

$$\mathcal{L}_\theta = \|\epsilon_\theta(\boldsymbol{x}_t, t, \boldsymbol{p}; \theta) - \hat{\boldsymbol{x}}_{\mathbf{0}}\|_2^2 . \tag{2}$$

Fine-grained Textual Control Signal $\boldsymbol{c}$ specifies the desired movements of particular body parts within defined temporal intervals. In this paper, we divide the human body into six main parts: *head*, *body*, *left arm*, *right arm*, *left leg*, and *right leg*. See the Appendix for details. An example of a fine-grained control signal for the *right leg* is shown in Fig. 3. Here, <SEP> separates the descriptions across different temporal intervals. <Motionless> indicates that no movement of the specified body part occurs in that interval.

Move your right leg forward. <SEP> Move your right leg back. Point your right foot on the floor. <SEP> Move your right leg forward. Bend your right knee. <SEP> <Motionless> <SEP> Move your right leg back.

Figure 3: A fine-grained textual control signal example.

3.3 Motion Generation with Fine-grained Textual Control Signal

To generate human motions that follow coarse-grained texts and the fine-grained textual control signals, we follow the ControlNet (Zhang, Rao, and Agrawala 2023) paradigm to propose a dual-branch framework, as illustrated in Fig. 2. The **lower branch** reuses the original MDM (Tevet et al. 2023) transformer encoder with resumed pretrained weights to ensure stable generation from coarse-grained texts $\boldsymbol{p}$. Specifically, the "*Input Process*" block includes a CLIP (Radford et al. 2021) text encoder to extract the textual embedding $\boldsymbol{e}_p$, and linear layers to encode the noisy motion sequence $\boldsymbol{x}_{\boldsymbol{t}}$ into a motion embedding $\boldsymbol{e}_{x_t}$. These two embeddings are then concatenated as

$$\boldsymbol{e} = [\boldsymbol{e}_p; \boldsymbol{e}_{x_t}], \tag{3}$$

and fed into the transformer blocks. The output of the l-th transformer block is defined as

$$\boldsymbol{h}_l^{\text{ori}} = \text{TransformerBlock}_l(\boldsymbol{h}_{l-1}^{\text{ori}}), \quad \text{with } \boldsymbol{h}_0^{\text{ori}} = \boldsymbol{e}. \tag{4}$$

Meanwhile, the **upper branch** is a trainable copy of MDM, modulated by the fine-grained textual control signal $\boldsymbol{c}$ through conditional feature adaptation. First, we construct the concatenated textual and motion embedding $\boldsymbol{e}'$ in the same way as the lower branch. To encode the control signal $\boldsymbol{c}$, we apply random masking on it by replacing descriptions within random temporal intervals with the special token <Mask>, and extract its embedding $\boldsymbol{e}_c$ using our well-designed text encoder (see Sec. 3.4). We then align $\boldsymbol{e}_c$ to the transformer's embedding space via a linear layer and add it to $\boldsymbol{e}'$, effectively modulating $\boldsymbol{e}'$ before it is fed into the transformer. The output of the l-th transformer block in the upper branch is computed as:

$$\boldsymbol{h}_l^{\text{ctrl}} = \text{TransformerBlock}_l(\boldsymbol{h}_{l-1}^{\text{ctrl}}), \tag{5}$$

$$\text{with } \boldsymbol{h}_0^{\text{ctrl}} = \boldsymbol{e}' + \text{Linear}(\boldsymbol{e}_c). \tag{6}$$

The upper branch is connected to the lower branch through linear layers $\mathcal{P}_l(\cdot)$ whose weights and biases are initialized to zero. Formally, the interaction between branches at the l-th layer is defined as:

$$\boldsymbol{h}_l^{\text{out}} = \boldsymbol{h}_l^{\text{ori}} + \mathcal{P}_l(\boldsymbol{h}_l^{\text{ctrl}}). \tag{7}$$

This zero-initialization prevents the injection of random noise into the lower branch during the early stages of training. As training progresses, the upper branch gradually learns to interpret the fine-grained textual control signals, capturing both spatial and temporal aspects, and refines the motion by injecting meaningful corrections into the corresponding layers of the motion diffusion model, thereby enhancing motion quality.

3.4 Hierarchical Contrastive Learning for Fine-Grained Textual Control Signal

Commonly used text encoders, such as CLIP (Radford et al. 2021) and T5 (Raffel et al. 2020), exhibit limited capability in extracting discriminative embeddings for fine-grained textual descriptions, especially our fine-grained textual control signals. This limitation hinders the performance of controllable motion generation. To solve this, we analyze the structure of our control signal and identify that it inherently contains three levels of information. As illustrated in Fig. 3:

- The whole example describes a specific body part's movements across all temporal intervals. Such information is referred to as **sequence-level**.
- Descriptions of a single interval, *i.e.*, between two <SEP> tokens, are considered as **snippet-level**.
- Each individual sentence within an interval is defined as **sentence-level**.

Building on this insight, we propose a hierarchical contrastive learning module with level-specific data augmentations to enhance the control signal embeddings. Specifically, we adopt T5 as the base text encoder due to its capacity to encode long text sequences, and progressively train it through contrastive learning at the sentence, snippet, and sequence levels, each initialized from the weights learned at the previous level. Next, we describe data augmentations for each level and the training objective in detail.

Sentence-Level. At this level, the goal is to enable the text encoder to distinguish between different body part movement sentences. Specifically, we build a sentence-level corpus $\mathcal{D}_{\text{sen}}^{\text{ori}}$ with non-repetitive body part movement sentences from FineMotion (Wu et al. 2025), which is the source of our fine-grained textual control signals. Then, we utilize DeepSeek-V2 (DeepSeek-AI 2024) to rewrite each sentence, forming an augmented corpus $\mathcal{D}_{\text{sen}}^{\text{aug}}$ (see Appendix for

prompts). Each sentence in $\mathcal{D}_{\text{sen}}^{\text{ori}}$ and its rewritten counterpart in $\mathcal{D}_{\text{sen}}^{\text{aug}}$ form a positive pair, while all other sentences in the corpora serve as their negative samples.

Snippet-Level. This level aims to make the text encoder robust to the order of sentences within a single time interval. Similarly, we collect snippet-level descriptions from FineMotion to build a snippet-level corpus $\mathcal{D}_{\text{sni}}^{\text{ori}}$. For augmentation, we randomly replace some of the sentences with their counterparts from $\mathcal{D}_{\text{sen}}^{\text{ori}}$ and shuffle their order (see Appendix, Algorithm 1). Each snippet-level description is augmented twice to form a positive pair; augmentations from other snippet-level ones serve as their negative samples.

Sequence-Level. The text encoder in this level aims to enhance its temporal awareness of different temporal intervals' body part movement descriptions, *i.e.*, our control signals. To construct augmented data at this level, we randomly apply snippet-level augmentations to the individual intervals within a control signal, but preserving the temporal order of intervals (see Appendix, Algorithm 2). Each sequence-level description is augmented twice to form a positive pair, while augmented versions of other sequence-level descriptions are treated as their negative samples.

Contrastive Learning. For all three levels, we train the text encoder with the InfoNCE (van den Oord, Li, and Vinyals 2019) loss, which pulls positive pairs closer and pushes negatives apart. In each minibatch, we sample N original texts and construct positive pairs $\boldsymbol{c}_i^{\text{aug}}$ and $\boldsymbol{c}_j^{\text{aug}}$ for each. Their embeddings can be represented as:

$$h_i = \text{Avg}(\mathcal{E}\left(\boldsymbol{c}_i^{\text{aug}}\right)), \quad h_j = \text{Avg}(\mathcal{E}\left(\boldsymbol{c}_j^{\text{aug}}\right)), \tag{8}$$

where $\mathcal{E}(\cdot)$ is the text encoder, and $\text{Avg}(\cdot)$ denotes the average pooling operation that aggregates the encoder outputs along the sequence length to produce a single embedding. A MLP projection head $g(\cdot)$, randomly initialized for each level, then maps the embeddings into contrastive space:

$$z_i = g(h_i), \quad z_j = g(h_j). \tag{9}$$

Thus, each minibatch yields $2N$ contrastive embeddings in total. Each embedding is contrasted against the other $2(N-1)$ negative ones. The contrastive loss for each embedding, such as $\boldsymbol{c}_i^{\text{aug}}$, is:

$$\mathcal{L}_i = -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right)/\tau\right)}, \tag{10}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, $\mathbb{1}(\cdot)$ is an indicator function, and τ is the temperature parameter.

4 Experiment

4.1 Experimental Setup

Datasets. We conduct experiments on the widely used HumanML3D (Guo et al. 2022a) dataset, using its texts as the coarse-grained texts $\boldsymbol{p}$. For fine-grained textual control signals $\boldsymbol{c}$, we use annotations from FineMotion (Wu et al. 2025), which provide detailed body part movement descriptions over short temporal intervals for HumanML3D motions. See the Appendix for details.

Implementation Details. We first progressively train the text encoder across three levels in the hierarchical contrastive learning module, then freeze it to extract control signal embeddings for training our FineXtrol framework. All experiments are conducted on a single A100 40G GPU. We follow the training hyperparameters from (Xie et al. 2024). See the Appendix for detailed hyperparameters.

Evaluation Metrics. The evaluation metrics are categorized into two groups: (1) Generated Motion Quality: Following (Xie et al. 2024), we assess the realism and diversity of generated motions using Frechet Inception Distance (FID), Multi-modal Distance (MM-Dist), R-Precision (Top-1/2/3 motion-to-text retrieval accuracy), and Diversity. For metric definitions, see (Guo et al. 2022a). (2) Textual Representation Quality: To evaluate the discriminability of embeddings for our fine-grained textual control signals, we adopt metrics from (Devlin et al. 2019; Gao, Yao, and Chen 2021), including cosine similarity, alignment, and uniformity. See their papers for details.

4.2 Comparison with State-of-the-arts

We compare different controllable motion generation methods on HumanML3D in Tab. 1. Results in blocks (1)-(3) are taken from (Xie et al. 2024; Wang et al. 2024), while others are our implementations. The implementation details can be seen in the Appendix. From Tab. 1, FineXtrol (*Body Part: Average*) achieves an FID of 0.245 (*vs.* 0.544) and an R-Top3 of 0.685 (*vs.* 0.611), compared to MDM without control signals in block (2), which demonstrates **the effectiveness of our novel control framework**. Compared to previous controllable motion generation methods in blocks (4)-(5), FineXtrol achieves state-of-the-art results in R-precision and Diversity, indicating its **ability to generate more diverse and accurate motions**. This supports the advantage of using our precise textual control signals, which offer greater flexibility than rigid spatial coordinates. Moreover, FineXtrol offers a more **user-friendly** approach to guide motion generation. We also compare results on controlling multiple body parts, a more challenging task, in block (6). Specifically, we follow OmniControl to train FineXtrol with combinations of body part control signals, resulting in a total of *i.e.*, $\binom{6}{1}+\binom{6}{2}+\binom{6}{3}+\binom{6}{4}+\binom{6}{5}+\binom{6}{6} = 63$ combinations. During evaluation, one combination is randomly sampled per

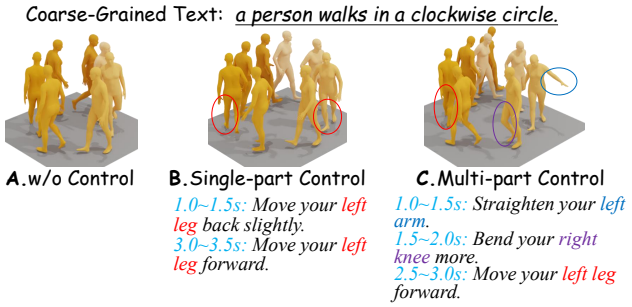

Figure 4: Visualizations of different control settings. <Mask> is used for all unspecified temporal intervals.

No.	Method	Control signal	User-Friendly control	Body Part	FID ↓	R-precision ↑ (Top-3)	Diversity →	MM-Dist ↓
(1)	Real	-	-	-	0.002	0.796	9.503	2.965
(2)	MDM (Tevet et al. 2023) ICLR'23	-	-	-	0.544	0.611	9.559	5.432
(3)	MDM (Tevet et al. 2023) ICLR'23	Coordinate	×	Pelvis Only	0.698	0.602	9.197	5.430
	PriorMDM (Shafir et al. 2024) ICLR'24		×		0.475	0.583	9.156	5.424
	GMD (Karunratanakul et al. 2023) CVPR'23		×		0.576	0.665	9.206	5.430
	OmniControl (Xie et al. 2024) ICLR'24		×		0.218	0.687	9.422	4.991
	InterControl (Wang et al. 2024) NeurIPS '24		×		0.159	0.671	9.482	5.026
(4)	InterControl (Wang et al. 2024) † NeurIPS '24	Coordinate	×	Average	**0.209**	**0.684**	**9.301**	5.164
	OmniControl (Xie et al. 2024) † ICLR'24		×		0.255	0.680	9.735	**5.054**
(5)	CoMo (Huang et al. 2024) † ECCV'24	Text	✓	Average	0.347	0.625	9.568	5.588
	FineXtrol (Ours)		✓		**0.245**	**0.685**	**9.492**	**5.087**
(6)	OmniControl (Xie et al. 2024) † ICLR'24	Coordinate	×	Cross	0.624	0.601	9.334	5.252
	CoMo (Huang et al. 2024)	Text	✓		0.606	0.611	9.662	5.638
	FineXtrol (Ours)	Text	✓		**0.351**	**0.676**	**9.658**	**5.146**

Table 1: Quantitative results on the HumanML3D test set. Methods in block (3) are trained with pelvis-only control, while those in blocks (4)-(6) are trained with control over all body parts. *Body Part (Average)* reports the average performance across all body parts under four levels of control signal density: 25%, 50%, 75%, and 100%. *Body Part (Cross)* reports the average performance over the combination of multiple body parts. † refers to our re-evaluation. → means closer to real data is better. Results show that our FineXtrol is efficient, user-friendly, and competitive with the existing controllable generation methods. We bold the best results in each block with body part: *Average* and *Cross*.

Body Part	FID ↓	R-Top3 ↑	Diversity →	MM-Dist ↓
Head	0.265	0.687	9.423	5.051
Body	0.261	0.679	9.548	5.119
Left Arm	0.224	0.684	9.671	4.981
Right Arm	0.219	0.684	9.427	5.208
Left Leg	0.239	0.685	9.543	5.063
Right Leg	0.263	0.689	9.341	5.100
Average	**0.245**	**0.685**	**9.492**	**5.087**

Table 2: Detailed results of controlling specific body parts.

test instance. Results show that other methods' *Cross* performance drops significantly, while FineXtrol experiences only a slight decline, highlighting the robustness of our approach.

Following (Xie et al. 2024; Wang et al. 2024), we display the detailed control performance (averaged over all density levels) for six different body parts in Tab. 2. The results indicate that using fine-grained textual control signal of *Right Arm* yields the best FID of 0.219, using those of *Right leg* achieves the highest R-Top3 of 0.689, and using those of *Left Arm* results in the best MM-Dist of 4.981. Detailed results of different control signal densities are provided in the Appendix. We also display our qualitative results in Fig. 4. It shows that FineXtrol supports both single- and multi-part control over different temporal intervals.

Furthermore, we report the inference time and the number of trainable parameters for our FineXtrol and other diffusion-based controllable motion generation models in Tab. 3. Specifically, we calculated the average time required to generate a single motion sequence with 1000 denoising steps, computed over 100 runs on an NVIDIA A100-SMX-40G GPU. The results show that FineXtrol has fewest trainable parameters, and achieves **highest inference efficiency**, benefiting from the absence of conversions between different pose representations.

Methods	Control Signal	Inference Time(s) ↓	Params.
OmniControl (Xie et al. 2024)	Coord.	168.51	48.79M
InterControl (Wang et al. 2024)		159.72	42.00M
GMD (Karunratanakul et al. 2023)		153.25	238.63M
FineXtrol (Ours)	Text	**128.57**	**23.39M**

Table 3: Inference time and the number of trainable parameters of diffusion-based controllable motion generation methods. 'Coord.' is short for coordinate.

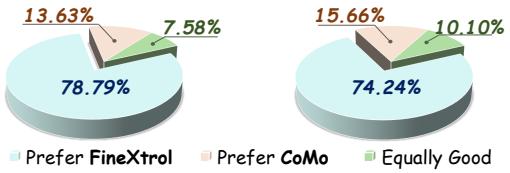

Figure 5: The statistical results of the user study. The *left* pie chart displays the average preference ratio for the visualized motion sequences without fine-grained textual control signals (2 cases) of our FineXtrol and CoMo. The *right* one shows that with fine-grained textual control signals (6 cases). Each case is evaluated based on (1) alignment with control signals and (2) motion naturalness.

We also conducted a user study with 33 subjects to compare our method with CoMo, which also uses fine-grained texts to enhance motions. Each subject evaluated 8 cases from two perspectives. According to the statistical results in

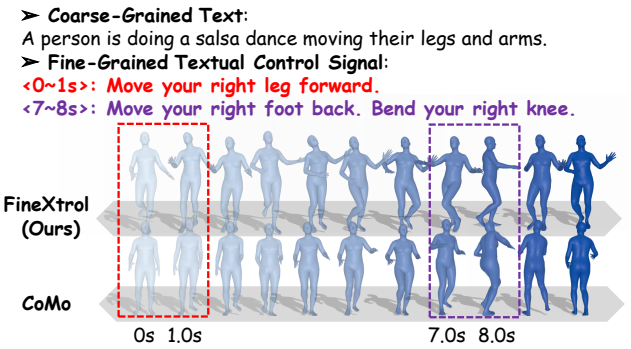

Figure 6: A motion pair comparing *right leg* control in the user study. Body part movements in unspecified intervals are not explicitly controlled.

the *left* part of Fig. 5, our method was preferred over CoMo in 78.79% of the cases without control signals. With control signals (*right* part of Fig. 5), 74.24% of users favored our results. Fig.6 illustrates a representative example: CoMo produces motions that are misaligned with control signals, whereas FineXtrol yields more faithful and precise motion. More comparisons are provided in the Appendix.

4.3 Ablation Study

We conduct ablation experiments on both the controllable motion generation paradigm and the textual embeddings to validate the effectiveness of our framework's design choices. We summarize key findings below.

Fine-grained Textual Control Paradigm	FID ↓	R-Top3 ↑
Direct	1.383	0.601
Ours	**0.245**	**0.685**

Table 4: Ablation study on control paradigm. Our control paradigm significantly outperforms the 'Direct' paradigm, which directly connects coarse-grained text and fine-grained textual control signals as a single input.

Effectiveness of Our Control Paradigm. We directly connect the fine-grained textual control signals with the coarse-grained text, encode the combined input using our text encoder, following Fig. 1(B), and denote this baseline as 'Direct'. Tab. 4 indicates that 'Direct' control paradigm yields poorer performance, even when the fine-grained descriptions are precise and temporally aware. This suggests that a single-branch model struggles to process densely packed information, and lacks the capacity to fully capture detailed textual semantics. In contrast, our paradigm not only preserves the ability to follow coarse-grained texts but also effectively guides the model to align with fine-grained textual control signals.

Effectiveness of the Hierarchical Contrastive Module for Fine-Grained Textual Control. We compare different text encoders by extracting embeddings for all fine-grained textual descriptions in FineMotion in Fig. 7. Results show

Text Encoder	FID ↓	R-Top3 ↑	Diversity →	MM-Dist ↓
CLIP	0.579	0.603	9.310	5.927
T5	0.374	0.659	9.594	5.483
Ours	**0.245**	**0.685**	**9.492**	**5.087**

Table 5: Ablation study on different text encoders. The controllable motion generation performance with our text encoder significantly surpasses those with CLIP and T5, proving the effectiveness of our hierarchical contrastive learning module for fine-grained textual control signals.

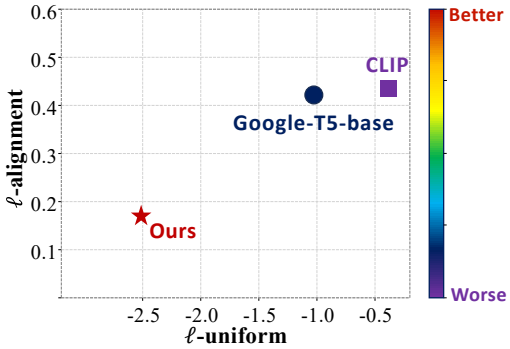

Figure 7: Comparison with existing text encoders in representing fine-grained textual control signals. We plot $\ell_{\text{align}} - \ell_{\text{uniform}}$ of textual embeddings from different text encoders. For both ℓ_{align} and ℓ_{uniform}, lower numbers are better. Results show that our text encoder trained with the proposed hierarchical contrastive learning module performs significantly better than prevalent text encoders.

that our text encoder, trained with the proposed hierarchical contrastive module, yields more discriminative embeddings than prior encoders. More ablations for each contrastive level are provided in the Appendix. We also assess their impact on fine-grained textual controllable motion generation in Tab. 5. Results show that the T5 encoder outperforms CLIP, likely due to CLIP's truncation of long, detailed texts. With our hierarchical contrastive module, our text encoder achieves further gains, surpassing both T5 and CLIP.

5 Conclusion

We propose FineXtrol, an efficient framework that leverages novel and user-friendly control signals, *i.e.*, precise and fine-grained textual descriptions of body part movements, to generate human motion sequences, offering a fresh perspective on controllable motion generation. We further introduce a hierarchical contrastive learning module to enhance the text encoder's ability to extract discriminative embedding for this novel control signal. Experimental results show that FineXtrol excels in controllable motion generation, particularly in controlling body part combinations. Visualizations further demonstrate its ability to control specific body parts, adjust movements within specific intervals, and manipulate multiple body parts via fine-grained textual descriptions.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2024YFF0618403), National Natural Science Foundation of China under Grant 62576216, and Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

References

Carlsson, F.; Gyllensten, A. C.; Gogoulou, E.; Hellqvist, E. Y.; and Sahlgren, M. 2021. Semantic Re-tuning with Contrastive Tension. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *CVPR*.

DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 6894–6910.

Giorgi, J. M.; Nitski, O.; Wang, B.; and Bader, G. D. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 879–895.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024. MoMask: Generative Masked Modeling of 3D Human Motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 1900–1910.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating diverse and natural 3d human motions from text. In *CVPR*.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*.

Harvey, F. G.; Yurick, M.; Nowrouzezahrai, D.; and Pal, C. 2020. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huang, Y.; Wan, W.; Yang, Y.; Callison-Burch, C.; Yatskar, M.; and Liu, L. 2024. CoMo: Controllable Motion Generation Through Language Guided Pose Code Editing. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIX*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Juravsky, J.; Guo, Y.; Fidler, S.; and Peng, X. B. 2022. PADL: Language-Directed Physics-Based Character Control. In *SIGGRAPH Asia 2022 Conference Papers*.

Kalakonda, S. S.; Maheshwari, S.; and Sarvadevabhatla, R. K. 2023. Action-GPT: Leveraging Large-scale Language Models for Improved and Generalized Action Generation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*.

Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *CVPR*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*.

Li, Y.; Hou, X.; Dezhi, Z.; Shen, L.; and Zhao, Z. 2024. FLIP-80M: 80 Million Visual-Linguistic Pairs for Facial Language-Image Pre-Training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 58–67. Association for Computing Machinery. ISBN 9798400706868.

Logeswaran, L.; and Lee, H. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2018. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 528–540.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 8748–8763.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 8821–8831.

Rempe, D.; Luo, Z.; Peng, X. B.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; and Litany, O. 2023. Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion. In *CVPR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Shafir, Y.; Tevet, G.; Kapon, R.; and Bermano, A. H. 2024. Human Motion Diffusion as a Generative Prior. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Starke, S.; Mason, I.; and Komura, T. 2022. DeepPhase: periodic autoencoders for learning motion phase manifolds. *ACM Trans. Graph.*

Tan, X.; Wang, H.; Geng, X.; and Zhou, P. 2024. Sopo: Text-to-motion generation using semi-online preference optimization. *arXiv preprint arXiv:2412.05095*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *ICLR*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Wang, H.; Weng, W.; Wang, J.; Zhao, F.; Xie, G.-S.; Geng, X.; and Wang, L. 2025a. Foundation model for skeleton-based human action understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; and Wei, F. 2022. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. *CoRR*.

Wang, X.; Hou, X.; Ding, M.; Chen, J.; Deng, K.; Xie, J.; and Shen, L. 2025b. DisFaceRep: Representation Disentanglement for Co-occurring Facial Components in Weakly Supervised Face Parsing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 4020–4029. Association for Computing Machinery.

Wang, Y.; Li, M.; Liu, J.; Leng, Z.; Li, F. W.; Zhang, Z.; and Liang, X. 2025c. Fg-T2M++: LLMs-augmented fine-grained text driven human motion generation. *International Journal of Computer Vision*, 1–17.

Wang, Z.; Wang, J.; Li, Y.; Lin, D.; and Dai, B. 2024. InterControl: Zero-shot Human Interaction Generation by Controlling Every Joint. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Wu, B.; Xie, J.; Ding, M.; Kong, Z.; Ren, J.; Bai, R.; Qu, R.; and Shen, L. 2025. FineMotion: A Dataset and Benchmark with Both Spatial and Temporal Annotation for Fine-Grained Motion Generation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wu, X.; Lai, Z.; Zhou, J.; Hou, X.; Pedrycz, W.; and Shen, L. 2024. Light-aware contrastive learning for low-light image enhancement. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9): 1–24.

Xie, Y.; Jampani, V.; Zhong, L.; Sun, D.; and Jiang, H. 2024. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Yan, S.; Li, Z.; Xiong, Y.; Yan, H.; and Lin, D. 2019. Convolutional sequence generation for skeleton-based action synthesis. In *ICCV*.

Yuan, Y.; Song, J.; Iqbal, U.; Vahdat, A.; and Kautz, J. 2023. Physdiff: Physics-guided human motion diffusion model. In *ICCV*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024. MotionDiffuse: Text-Driven Human Motion Generation With Diffusion Model. *IEEE Trans. Pattern Anal. Mach. Intell.*