Noname manuscript No.
(will be inserted by the editor)

Fg-T2M++: LLMs-Augmented Fine-Grained Text Driven Human Motion Generation

Yin Wang · Mu Li · Jiapeng Liu · Zhiying Leng · Frederick W. B. Li · Ziyao Zhang · Xiaohui Liang ✉

Received: date / Accepted: date

Abstract We address the challenging problem of fine-grained text-driven human motion generation. Existing works generate imprecise motions that fail to accurately capture relationships specified in text due to: (1) lack of effective text parsing for detailed semantic cues regarding body parts, (2) not fully modeling linguistic structures between words to comprehend text comprehensively. To tackle these limitations, we propose a novel fine-grained framework Fg-T2M++ that consists of: (1) an *LLMs semantic parsing module* to extract body part descriptions and semantics from text, (2) a *hyperbolic text representation module* to encode relational information between text units by embedding the syntactic dependency graph into hyperbolic space, and (3) a *multi-modal fusion module* to hierarchically fuse text and motion features. Extensive experiments on HumanML3D and KIT-ML datasets demonstrate that Fg-T2M++ outperforms SOTA methods, validating its ability to accurately generate motions adhering to comprehensive text semantics.

Keywords Text Driven Motion Generation · Human Motion · Diffusion Model · Large Language Model

Yin Wang
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
E-mail: wang_yin@buaa.edu.cn

Mu Li
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
E-mail: limu@buaa.edu.cn

Jiapeng Liu
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
E-mail: zy2306414@buaa.edu.cn

Zhiying Leng
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
E-mail: zhiyingleng@buaa.edu.cn

Frederick W. B. Li
Department of Computer Science, University of Durham, U.K
E-mail: frederick.li@durham.ac.uk

Ziyao Zhang
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
E-mail: 20373042@buaa.edu.cn

Xiaohui Liang (✉Corresponding author)
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
Zhongguancun Laboratory, Beijing, China
E-mail: liang_xiaohui@buaa.edu.cn

1 Introduction

Human motion generation is a pivotal but challenging task in computer vision with applications in animation, AR/VR, gaming etc. While existing works utilize diverse multimodal inputs such as music (Kao and Su, 2020; Li et al., 2021; Ren et al., 2020; Starke et al., 2022; Tseng et al., 2023), motion categories (Guo et al., 2020; Petrovich et al., 2021; Cervantes et al., 2022; Guo et al., 2022c) and trajectories (Karunratanakul et al., 2023; Shafir et al., 2023; Wan et al., 2023), collecting and annotating such data requires substantial expertise. Recently, text-driven human motion generation (T2M) has emerged as a promising research direction, which generates motions represented by 3D joint positions or rotations (Guo et al., 2022a; Petrovich et al., 2022; Tevet et al., 2023; Chen et al., 2023; Zhang et al., 2024a). T2M holds potential for intuitive motion control through semantics-rich text inputs. However, key challenges remain in parsing intricate spatiotemporal relationships between body parts from descriptions and

Fig. 1: Our Fg-T2M++ excels in generating high-quality and diverse motion sequences, capturing fine-grained details embedded in the text prompts.

mapping diverse linguistic expressions to realistic motions. While promising an accessible interface, addressing these issues is non-trivial. Advances in T2M could unlock applications from animation to assistive technologies through natural guidance.

Three main approaches are proposed to address text-driven human motion generation: (1) Latent space alignment methods, such as JL2P (Ahuja and Morency, 2019) and TEMOS (Petrovich et al., 2022), aim to learn a shared latent space between text and motion representations by directly integrating their embeddings. However, this integration can potentially lead to the loss of modality-specific information. (2) Conditional autoregressive models generate motion tokens sequentially, conditioned on previous tokens and text. Pioneering works like TM2T (Guo et al., 2022b) employ vector quantized VAEs to decode motion tokens from discrete representations learned from data, while T2M-GPT (Zhang et al., 2023a) enhances this with techniques such as exponential moving average and code resetting for more natural generation. Despite their strengths in capturing temporal dependencies, these methods rely on unidirectional and sequential prediction, which can impact the quality of motion generation due to cumulative errors. (3) Conditional diffusion models, including MotionDiffuse (Zhang et al., 2024a) and MDM (Tevet et al., 2023), adopt diffusion frameworks to probabilistically map text to motion via denoising training objectives, achieving promising performance. The condition plays a crucial role in guiding the denoising process. However, current methods often lack refinement in handling these conditions. Firstly, the exploration of additional auxiliary information is insufficient because datasets like HumanML3D (Guo et al., 2022a) and KIT-ML (Plappert et al., 2016) offer only coarse descriptions without fine-grained part-level annotations. Secondly, the extraction and fusion of conditional features are inadequate; current methods typically extract compact sentence representations from text, failing to fully utilize rich information within words. This limitation can result in generated motions deviating from the original text meaning when simply concatenating sentence-motion vectors.

Existing text-to-motion methods struggle with generating whole-body motions for unseen text, as full motion may lie outside the training distribution, yet individual body part motions still fall within it. We address this by hypothesizing decomposition of whole-body generation into combinable sub-joint motions of multiple parts facilitates easier modeling. Also, natural language semantically encodes actions through parts-of-speech and syntactically relates words through grammatical structures. To capture fine-grained linguistic details, we propose analyzing individual body part motions specified by words, considering their syntactic roles and relationships. Based on these insights, we introduce the novel framework Fg-T2M++ (Figure 1), leveraging part-level and word-level natural language descriptions to generate precise motions conditioned on text prompts. This approach decomposes generation and deeply analyzes textual details, aiming to overcome limitations in generating whole-body and fine-grained motions.

Fg-T2M++ comprises three integrated components for fine-grained language-guided motion generation, each serving distinct purposes. First, our *LLMs Se-*

mantic Parsing (LSP) module uses large language models to extract detailed semantic descriptions from text prompts. It parses the text into annotations of individual body part motions and their relationships through deep linguistic analysis of semantic roles between parts of speech (e.g., nouns, adjectives) and motions. This fine-grained parsing maps individual textual elements to joint movements, allowing for the understanding of complex language beyond prior methods that relied on shallow encodings. Second, the *Hyperbolic Text Representation (HTP) module* focuses on encoding the syntactic structure of text prompts by constructing a dependency parse tree and embedding it in hyperbolic space. Hyperbolic geometry intrinsically preserves hierarchies with low distortion (Yang et al., 2022), enabling HTP to capture hierarchical relationships more effectively than Euclidean models. Third, to achieve a fine-grained fusion of multimodal information, our *Multi-Modal Fusion (MMF) module* hierarchically fuses outputs from the HTP and LSP modules at both global and local levels. It combines global and local features to learn comprehensive text-motion mappings. This multi-scale fusion of syntactic and semantic information provides a comprehensive understanding of text-motion mappings not achieved by prior global or local modeling alone. By integrating linguistically-informed parsing, hyperbolic syntactic modeling, and hierarchical semantic fusion, Fg-T2M++ captures fine-grained text-motion correlations in a technically advanced yet concise manner compared to prior works.

While our previous work, Fg-T2M, in ICCV 2023 (Wang et al., 2023), made progress in text-driven motion generation, it was incapable of addressing the novel research problems associated with capturing fine-grained motion details specified in text. This limitation arose from its coarse-grained modeling of syntactic relationships without a detailed analysis of text prompts. To address this, we propose Fg-T2M++ with novel technical contributions - LLMs Semantic Parsing to extract body part-level semantics from text, hyperbolic text representation module encoding hierarchical dependency graphs in hyperbolic space, and multi-modal fusion performing multi-level fusion within a conditional diffusion framework. This design further extends Fg-T2M's capabilities for fine-grained tasks. Extensive evaluation demonstrates the effectiveness of Fg-T2M++ over Fg-T2M, achieving a significantly lower FID of 0.135 versus 0.571 and MM-Dist of 2.696 versus 3.114 on KIT-ML, validating its ability to generate motions specified by richer textual details that Fg-T2M was technically unable to capture. Our main contributions are:

- We propose an LLMs Semantic Parsing module to parse text into fine-grained body part representations and detailed words semantics leveraging large language models.
- We introduce a Hyperbolic Text Representation module incorporating dependency parsing and hyperbolic graph convolution to embed syntactic trees in hyperbolic space, exploiting its advantages over Euclidean space.
- We present a Multi-Modal Fusion module performing hierarchical fusion of global and local text-motion relationships within a conditional diffusion framework through multiple denoising steps.
- We validate Fg-T2M++ on HumanML3D and KIT-ML datasets, demonstrating SOTA performance through metrics, and qualitative results revealing finer motion generation matching text.

2 Related Work

2.1 Text Driven Human Motion Generation

While latent space alignment works such as JL2P (Ahuja and Morency, 2019) and Ghosh et al. (Ghosh et al., 2021) achieved progress utilizing joint embeddings and hierarchical encoders capturing coarse relationships, as well as techniques like MotionCLIP (Tevet et al., 2022) that generates stylized motions by projecting into a shared space learned via CLIP (Radford et al., 2021), TEMOS (Petrovich et al., 2022) combining motion and text VAEs, and temporal VAE (Guo et al., 2022a) for sequence generation, their limitation is loss of fine-grained details when encoding independently.

Autoregressive models such as TM2T (Guo et al., 2022b), which learns mutual mappings of motion and tokens via vector quantized VAEs, T2M-GPT (Zhang et al., 2023a), which enhances performance with EMA and code resetting, and AttT2M (Zhong et al., 2023) mapping to refined codes via body part attention, have achieved progress in representing motion as discrete tokens. However, the unidirectional nature of autoregressive models limits their ability to capture future context, affecting motion quality. Incorporating bidirectional dependencies could improve this but increase training and inference costs due to the additional computational complexity involved.

Recent diffusion-based models, such as MotionDiffuse (Zhang et al., 2024a), MDM (Tevet et al., 2023), and FLAME (Kim et al., 2023), have shown promising performance in T2M tasks by leveraging conditional diffusion to learn probabilistic text-motion mappings. Also, MLD (Chen et al., 2023) employs latent diffusion to enhance efficiency, while ReMoDiffuse (Zhang et al.,

2023b) incorporates sample retrieval for contextual understanding. However, they may suffer from a lack of fidelity in generating motions that precisely align with conditional inputs, particularly in capturing complex multi-modal relationships.

Existing methods perform relatively well on coarse-grained text, such as "a person is walking." However, they struggle with fine-grained text that involves complex syntax-kinematic associations, like "a person is walking with the right hand raising while stumbling to the left." Latent space alignment methods can lose significant feature details during feature projection, often processing only the common "walking" motion. Autoregressive models, due to their unidirectional prediction nature, may overlook subsequent movements like "stumbling", leading to motion incoherence. Diffusion-based models face challenges in feature extraction and fusion between the denoising sequence and text, which can result in ignoring finer details such as "right hand raising" or "stumbling to the left."

Despite advances in text-to-motion generation, challenges remain regarding fine-grained modeling. The sparsity of current datasets limits learning precise textual cue-motion correspondences. Insufficient use of linguistic cues also restricts comprehending fine-grained semantics from prompts. Addressing these issues, we created detailed annotations of different body parts' actions and words explanations, enabling more intricate understanding of part-specific details. Furthermore, leveraging linguistic structures assists semantic parsing of prompts. This enables our model to generate human motions closely aligned with the semantic content of input text, exhibiting realistic movements.

2.2 LLMs-Assisted Motion Generation

While large language models such as BERT (Devlin et al., 2018), GPT-4 (Achiam et al., 2023) and T5 (Raffel et al., 2020) have demonstrated strong capabilities in language tasks as evidenced by their human-level performance in certain domains (Gilardi et al., 2023), their application to human motion generation has strengths and limitations. Recent works including ActionGPT (Kalakonda et al., 2023), SINC (Athanasiou et al., 2023), FineMoGen (Zhang et al., 2024b), and MotionGPT (Jiang et al., 2024) have explored leveraging LLMs' language generation and zero-shot transfer abilities to enrich prompts, identify body parts, facilitate human-AI interaction, and support various motion-related tasks. However, directly generating coherent human motions from language remains challenging due to the complex grounding problem between language and bodily motion. Their suitability ultimately depends on how effectively language representations can condition low-dimensional movement sequences.

Previous works utilizing LLMs have achieved good results, yet challenges for improving fine-grained analysis and modeling persist. Urgently needed are datasets with precise, fine-grained text representations that are sensitive to subtle motion details. To address this, we introduce the use of LLMs for parsing text prompts at a fine level to obtain specific descriptions of individual body parts. We also provide detailed explanations of nouns, adjectives, and adverbs in sentences to address challenging vocabulary in complex texts. By training models with these fine-grained linguistic details regarding all body parts and words, we can generate high-fidelity, fine-level human motion sequences.

3 Preliminaries

3.1 Diffusion Model

Our Fg-T2M++ model for fine-grained text-to-motion generation is based on the diffusion probabilistic framework. As described in prior work (Ho et al. (2020)), diffusion models comprise a forward noise injection process and reverse conditional generation process. In the forward process, a clean target motion sequence x_0 is gradually corrupted with added Gaussian noise to produce a simple Gaussian distribution. This defines an auto-encoding formulation.

Crucially, in the reverse process, noise is removed from the corrupted motion sequence $x_1, ..., x_T$ in a conditional manner given natural language text c. We model this conditional generation as:

$$p_\theta(x_{0:T}|c) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c) \tag{1}$$

By conditioning each step of the diffusion posterior on both the motion context and linguistic input c, our Fg-T2M++ model can progressively generate fine-grained target motions aligned precisely with the text description. This principled diffusion formulation underpins our ability to capture rich, detailed language-motion mappings.

3.2 Hyperbolic Graph Convolution

Hyperbolic space H^n is a non-Euclidean space with constant negative curvature, represented as an n-dimensional Riemannian manifold. Among models like the Poincaré ball, Lorentz, and Klein models, we use the Poincaré ball model $B_c^n = \{x \in \mathbb{R}^n \mid \|x\|^2 < r^2\}$ for

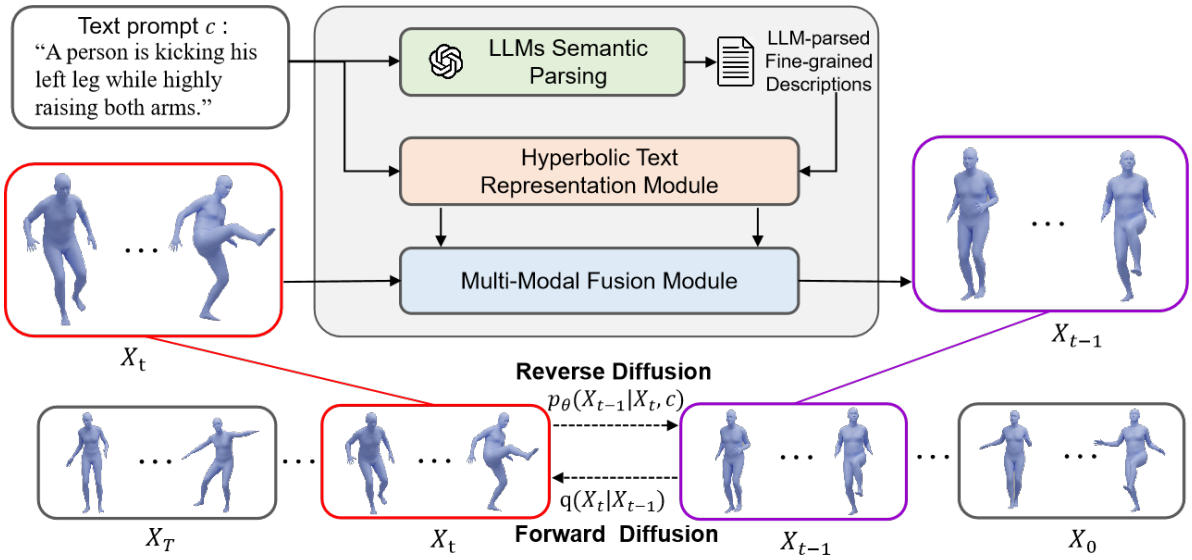

Fig. 2: **Overview of Fg-T2M++**: Given a text prompt c, the reverse denoising process of the diffusion model starts from noisy motion data X_T and produces clean motion data X_0. Initially, the text prompt undergoes LLMs semantic parsing to generate LLMs-parsed fine-grained descriptions. Then, both the text prompt and its parsed descriptions are input into the hyperbolic text representation module, which captures precise representations of text features. Finally, the noisy motion data X_t, along with the two fine-grained text features, are fed into the multi-modal fusion module to obtain the clean motion data X_{t-1}.

its geometric property preservation. Here, c is the curvature, while $r = \frac{1}{\sqrt{c}}$ defines the radius. The variable x denotes a position within the Poincaré ball, serving as the text embedding vector in our T2M generation model.

The Poincaré ball has Riemannian metric $g_x^H = (\lambda_x^c)^2 g^E$ conformal to the Euclidean metric g^E with $\lambda_x^c = \frac{2}{1-||x||^2}$. This exponential shrinks distances towards the boundary, allowing hierarchical structures to be represented with a large branching factor.

We employ the exponential map $\text{Exp}_x : \tau_x H \to H$ and its inverse, the logarithmic map Log_x, to project points between the hyperbolic Poincaré ball and Euclidean space, where $\tau_x H$ refers to the tangent space at a point x in the hyperbolic space H. This enables us to embed syntactic trees extracted from text as hyperbolic graphs. Hyperbolic space is well-suited for such hierarchical data due to its ability to model low-dimensional structures with minimal distortion compared to Euclidean counterparts (Nickel and Kiela, 2017; Leng et al., 2023). For linguistic trees in our task, it thus provides a more suitable geometric domain.

Hyperbolic Graph Convolution (HGC) provides a powerful way to learn representations of hierarchical graph-structured data like syntactic trees in our task. HGC generalizes graph convolutional operations to hyperbolic space by projecting node features from the ambient Euclidean space to the Poincaré ball using the exponential map $\text{Exp}(\cdot)$ (Liu et al., 2019). It then applies the $\mathcal{M}$öbius layer operations $\otimes$ and $\oplus$ to transform features while preserving distances between embedded points (Kochurov et al., 2020). Neighborhood aggregation is performed via hyperbolic pooling functions $\mathbf{F}^H$ to combine neighboring node representations (Liu et al., 2019). Finally, the hyperbolic activation σ^H introduces non-linearities by alternating between Riemannian and Euclidean spaces. Compared to Euclidean GNNs, HGC's ability to scale vectors proportional to their hyperbolic distance allows better embedding of trees with minimal distortion. This makes it critical for modeling syntactic dependencies in our text using hyperbolic graph embeddings $\mathbb{G} = (\nu, \xi)$, where ν denotes all the nodes, specifically referring to each word in a sentence, and ξ represents all the edges, specifically indicating the syntactic relationships between each pair of words. HGC thus provides a powerful way to learn task-specific representations of our hierarchical input data.

4 Methodology

Given a text prompt, $\mathbf{W} = \{w_1, w_2, \ldots, w_N\}$, $\mathbf{W} \in \mathbb{R}^{N \times L}$ where N represents the number of words and L is the dimension of word vector. Our goal is to generate a human motion sequence, denoted as $\mathbf{M} =$

You are an AI skilled in language analysis.
Now, I will provide some senteances, and you need to provide a general analysis of the motion, a detailed analysis, and identify the verbs, as well as nouns (other than 'person' and 'human'), adjectives, adverbs, quantifiers, and conjunctions in the sentence. Explain their meanings and how they relate to bodily movements. I will provide you with a list of body parts, dividing the human body into six parts.

In your analysis, use only these six parts to describe movements.

For verbs:
Provide a general and a detailed analysis, as per the example.

For nouns:
Analyze the nouns in the sentence. Since most of these sentences start with 'person,' which doesn't contribute much to the motion generation, focus on other nouns, which are mostly objects. Describe these nouns in detail and explain who they are acted upon.

For adjectives:
Analyze the adjectives in the sentences. Describe what these adjectives mean, which parts of the body list they involve, to what extent do these body parts move, and who are they acting upon?

For adverbs:
Analyze the adverbs in the sentences. Describe what these adverbs mean, which parts of the body list they involve, and who are they acting upon?

For quantifiers:
Analyze the quantifiers in the sentences. What do these quantifiers mean? Which parts of the body list do they control? Who are they acting upon?

For conjunctions:
Analyze the conjunctions in the sentence (excluding 'is,' 'are'). What do these conjunctions mean? Which parts of the body list do they control? What is the order of words controlled by these conjunctions?

Format:
{"v1": "General verb analysis", "v2": "Detailed verb analysis (left arm)", "v3": "Detailed verb analysis (right arm)", "v4":"Detailed verb analysis (left leg)", "v5": "Detailed verb analysis (right leg)", "v6":"Detailed verb analysis (head)", "v7": "Detailed verb analysis (torso)", "n1": "Noun analysis", "adj1": "Adjective analysis", "adj2": "Body parts involved for adjectives", "adv1": "Adverb analysis", "adv2": "Body parts involved for adverbs" "quant1": "Quantifier analysis", "quant2": "Body parts involved for quantifiers", "conj": "Conjunction analysis"}

Example1: A person walks 4 steps forward and waves his left hand ,then play with a guitar.
{"v1": "A person first walks 4 steps forward, then waves his left hand, and then plays the guitar with the left hand holding the neck and the right hand strumming the strings, maintaining body stability and a slight head tilt to watch the guitar.",
"v2": "Left Arm : Used for waving, then grips the guitar neck during play.",
"v3": "Right Arm: Natural during walking, becomes active during waving and strumming the guitar strings.",
"v4": "Left Leg : Alternates with the right leg during walking, remains stationary while playing the guitar.",
"v5": "Right Leg : Moves alternately with the left leg during walking, stays stationary during guitar play.",
"v6": "Head : Steady and facing forward during walking and hand waving, slightly tilts during guitar play for a better view of the fretboard and strings.",
"v7": "Torso: Maintains stability throughout, may lean forward slightly while playing the guitar.",
"n1": "guitar: A stringed musical instrument with six strings, comprising a neck, a resonant body, and strings, used for playing music. hands: A part of the body consisting of the palm and five fingers, connected to the forearm. step: A movement in walking, involving lifting the foot and putting it down in a new position, part of walking or moving forward.",
"adj1": "left: Refers to the side of the body closer to the heart.",
"adj2": "left hand is involved in the waving action.",
"adv1": "forward: Indicates a direction or position of advancement.",
"adv2": "Whole Body moves forward during walking.",
"quant1": " ",
"quant2": " ",
"conj": "and: Connects the actions of walking and waving. then: Indicates a sequence of waving followed by playing the guitar."}

Fig. 3: The prompt of strategy and example for LLMs Semantic Parsing.

$\{m_1, m_2, \dots m_S\}$, where $\mathbf{M} \in \mathbb{R}^{S \times D}$. Here, S refers to the motion sequence length and D is the pose representation dimension. To achieve this, we present a diffusion model-based framework, Fg-T2M++, for fine-grained and high-fidelity text-driven motion generation. In the following sections, we provide: an overview of our motion generation approach in Section 4.1; introduction to the LLMs Semantic Parsing in Section 4.2; the Hyperbolic Text Representation Module in Section 4.3; and the Multi-Modal Fusion Module in Section 4.4.

4.1 Motion Generation via Diffusion Models

Our Fg-T2M++ approach generates fine-grained motions conditioned on natural language using a diffusion probabilistic model. As shown in Figure 2, we sample random noise and input it to the diffusion model along with timestep T and text condition c. The model iterates backwards from X_T to X_0, removing noise at each step. Crucially, at each denoising step the text is first parsed by LLMs into fine-grained part-level descriptions. The HTP module then encodes the text using a hyperbolic linguistics tree representation. Finally, the MMF module collaboratively reasons over the noisy motion and rich text encodings to acquire the clean motion embedding.

We employ classifier-free diffusion guidance (Ho and Salimans, 2022) to scale conditional and unconditional distributions as:

$$\epsilon = s\epsilon_\theta(x_t, t, c) + (1 - s)\epsilon_\theta(x_t, t, \varnothing) \tag{2}$$

where guidance scale s controls the text conditioning. Our objective predicts the clean state X_0 by minimizing the L2 loss between predicted and ground truth motions, enabling Fg-T2M++ to learn high-quality generation via:

$$\mathcal{L} = \mathrm{E}[\| \mathbf{x}_0 - \epsilon_\theta(\mathbf{x}_t, t, c) \|_2^2] \tag{3}$$

where $\epsilon_\theta(\mathbf{x}_t, t, c)$ denotes the model predict output. This diffusion formulation empowers our approach for the challenging task.

4.2 LLMs Semantic Parsing

Existing datasets provide rich motion data but have coarse text prompts limited to brief action descriptions like “a person is jumping” (Plappert et al., 2016; Guo

et al., 2022a), which may overlook fine-grained information of other body parts, such as hand waving details. The absence of such detailed descriptions significantly hinders fine-grained motion generation. Prior work encodes only high-level semantics from coarse prompts using shallow representations. As a result, generated motions do not precisely match detailed language specifications, such as coordinated part-level movements over time. We aim to synthesize motion from fine-grained natural language descriptors. However, current methods cannot comprehend such rich descriptions.

Large language models have advanced NLP through powerful modeling, e.g. OpenAI's GPT (Achiam et al., 2023). Our LLMs Semantic Parsing approach leverages this by parsing prompts into annotations of part motions and semantics (e.g. nouns, adjectives, adverbs, etc.). This fine-grained parsing captures specifications enabling intricate linguistic-kinematic modeling. It closes the gap between coarse data and our goal of animating complex language descriptions through state-of-the-art techniques.

Our approach utilizes strong priors from LLMs like GPT-3.5 to precisely capture relationships between natural language and human motion at a fine-grained level. For the given text, we perform dual parsing of action and semantics. To parse action, we represent the human skeleton from SMPL (Loper et al., 2023) and MMM (Terlemez et al., 2014) as six main body parts - left arm, right arm, left leg, right leg, head, and torso. Leveraging GPT's understanding of language and motion knowledge, we split whole-body movements described in the text into sub-joint motions of individual parts. Inspired by ActionGPT (Kalakonda et al., 2023) and GraphMotion (Jin et al., 2024), verbs are further clarified due to their decisive role in sentences.

We also parse semantics by focusing on parts of speech like nouns, adjectives, adverbs, quantifiers, and conjunctions, which convey richer meanings than other words. For instance, we interpret how adjectives and adverbs modify specific joints and how conjunctions connect actions. In total, each prompt undergoes 15 sub-analyses providing comprehensive parsing of both action and semantics. As Figure 3 illustrates, this dual fine-grained strategy allows comprehending motion sequence details at a granular level from the text.

4.3 Hyperbolic Text Representation

Prior work in language-guided motion generation encodes text directly using Transformers (Vaswani et al., 2017; Zhang et al., 2024a; Tevet et al., 2023; Chen et al., 2023). While capturing high-level semantics, this approach struggles to model the fine-grained details

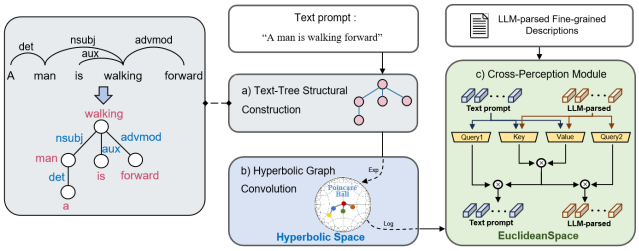

Fig. 4: **Architecture of HTP**. a): the process of text-tree structural construction by dependency analysis. b): the process of hyperbolic graph convolution in the hyperbolic space to grasp the texts' precise features. c): the process of cross-perception module to make full use of the LLMs-parsed fine-grained descriptions.

needed to comprehend language fully. However, text inherently contains a rich hierarchical structure, with word interconnections providing contextual information beyond standalone semantics. This structural information is not sufficiently utilized by existing methods. To address this, we propose the Hyperbolic Text Representation Module (HTP) to leverage the inherent syntactic structure of language prompts. HTP extracts and utilizes this structural knowledge via components like text-tree construction and hyperbolic graph networks. By incorporating the additional contextual cues provided by linguistic structure, HTP facilitates improved comprehension over prior semantic encoding-focused approaches. It aims to generate motion more precisely synchronized to prompt specifications.

Text-Tree Structural Construction. We leverage dependency parsing to identify syntactic relationships between phrases and address issues in prior work. Dependency parsing analyzes parts of speech and dependencies between words using spaCy, an NLP library for text processing functions like syntactic parsing (Honnibal and Montani, 2017). For a given prompt, dependency parsing establishes a text relationship tree where each word is a node and dependencies form connecting edges (Figure 4a). This tree serves as the initialization for a graph, providing prior knowledge of its topological structure compared to methods lacking syntactic context modeling. The tree represents phrases and their interdependencies more comprehensively than isolated words. This fine-grained structural information facilitates a deeper understanding of the text beyond singular semantics. By constructing linguistic trees from prompts, we aim to generate motion sequences synchronized more precisely to language descriptions.

Hyperbolic Graph Convolution. HGC is well-suited for processing tree-structured linguistic data in hyperbolic versus Euclidean space, as hyperbolic geom-

etry preserves local structure with low distortion (Yang et al., 2022). We construct a graph from the dependency parsed text tree, where words are nodes and dependencies are directed edges. Text embeddings $\mathbf{W}^E$ extracted from CLIP (Radford et al., 2021) and edge relationships ξ are combined to construct the graph $\mathbf{G}^E = \{\mathbf{W}^E, \xi\}$, where E represents Euclidean space. This graph is projected into the Poincaré ball hyperbolic model via the exponential function Exp: $\mathbf{G}^H = \text{Exp}(\mathbf{G}^E)$, where H represents hyperbolic space.

Within this hyperbolic manifold, stacked HGC layers process the graph through $\mathcal{M}$öbius calculus and hyperbolic nonlinear activations σ^{H}. This updates node features to capture hierarchical structure (Figure 4b). Features are then projected back to Euclidean space using the inverse exponential function Log:

$$\mathbf{W}^h = \text{Log}(\sigma^H(\mathcal{M}\text{öbius}(\text{Exp}(\mathbf{W}^E)))) \tag{4}$$

By applying HGC to model the linguistic structure, we obtain text encodings informed by both dependencies and syntax contexts.

Cross-Perception Module. In the parsing phase, our LLMs-Augmented approach precisely captures movement and linguistic details from prompts, allowing deep understanding. For parsed sentences, we use CLIP to obtain initial encodings and enhance them by passing through a Transformer layer, obtaining $\mathbf{W}^l$. For full text prompts, our HGC captures structural semantics at multiple levels, outputting encodings. These are also enhanced by a Transformer layer, resulting in $\mathbf{W}^t$.

The Cross-Perception Module aims to enhance text representations by modeling relationships between the parsed and text prompts encodings. As shown in Figure 4c, it applies a multi-stage attention process. Inspired by efficient attention (Shen et al., 2021), global context features $\mathbf{F}$ are computed using key-value attention over the concatenated encodings:

$$\mathbf{F} = \text{softmax}(\mathbf{Key}[\mathbf{W}^l; \mathbf{W}^t]) \otimes (\mathbf{Value}[\mathbf{W}^l; \mathbf{W}^t]), \tag{5}$$

where $[\cdot;\cdot]$ indicates a concatenation of two tensors. Cross-attention is then applied using query vectors $\mathbf{Q}^l$ and $\mathbf{Q}^t$:

$$\begin{aligned} \mathbf{W}^l &= \mathbf{W}^l + \text{softmax}(\mathbf{Q}^l\mathbf{W}^l) \otimes \mathbf{F}, \\ \mathbf{W}^t &= \mathbf{W}^t + \text{softmax}(\mathbf{Q}^t\mathbf{W}^t) \otimes \mathbf{F}. \end{aligned} \tag{6}$$

The enriched encodings capture multi-level relationships to guide motion generation.

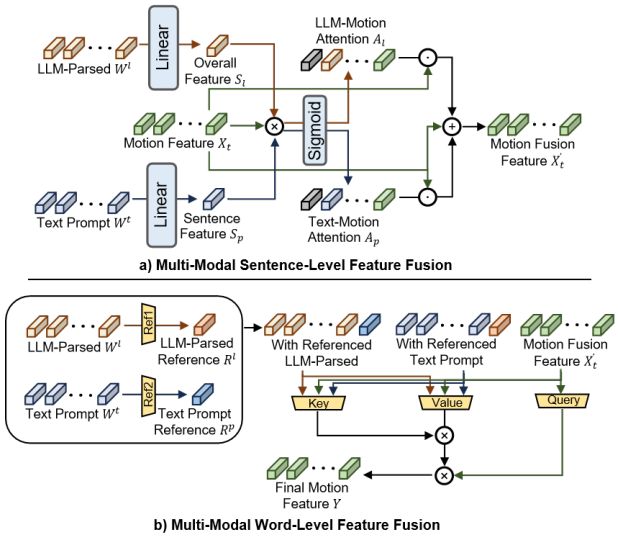

Fig. 5: **Illustration of two fusion methods in MMF.** a) multi-modal sentence-level feature fusion and b) multi-modal word-level feature fusion.

4.4 Multi-Modal Fusion

Existing methods that learn fixed word features struggle to capture high-order semantics (Tevet et al., 2023; Zhang et al., 2024a, 2023b). However, human sentence comprehension proceeds hierarchically from coarse to fine. To better model this, we introduce a coarse-to-fine structure in our motion diffusion model comprising two semantic levels: overall and detailed information.

Our proposed Multi-Modal Fusion (MMF) module aims to iteratively refine the interaction between text and motion encodings for controlling fine-grained motion diffusion. As shown in Figure 5, it contains two parts:

1. Multi-modal sentence-level feature fusion captures the overall semantic meaning across encoded text and motion modalities.
2. Multi-modal word-level feature fusion iteratively refines text-motion features through attention to a reference sequence, given the overall semantic context computed above.

Within our hierarchical Fg-T2M++ generation framework, MMF utilizes encoded representations from the Cross-Perception module. It provides progressively refined text signals to guide motion diffusion from coarse to fine-grained details. This hierarchical reasoning approach allows better modeling of language at both global and local levels, addressing challenges in prior work with fixed encodings.

Multi-Modal Sentence-Level Feature Fusion aims to combine multi-modal control signals from

parsed content and text prompts at the overall semantic level. As illustrated in Figure 5a, we first transform the parsed content into overall features S_l and the text prompts into sentence features S_p. We then calculate frame-level attention maps A_l and A_p denoting relevance between each motion feature X_t and the sentence features:

$$\mathbf{A}_l = \mathbf{X}_t(\mathbf{S}^l)^T, \mathbf{A}_p = \mathbf{X}_t(\mathbf{S}^p)^T \tag{7}$$

This captures correspondence between visual motion and linguistic semantics. The cross-modal motion features X'_t are obtained by highlighting channels in X_t related to both sentences:

$$\mathbf{X}'_t = \mathbf{X}_t + \lambda_l(\mathbf{X}_t \odot \sigma(\mathbf{A}_l)) + \lambda_p(\mathbf{X}_t \odot \sigma(\mathbf{A}_p)) \tag{8}$$

where λ_l and λ_p terms control contribution and σ is the sigmoid activation. This fusion operates at the coarse semantic level to provide context for the following word-level feature refinement.

Multi-Modal Word-Level Feature Fusion captures fine-grained text-motion relationships. Inspired by ReMoDiffuse (Zhang et al., 2023b), it iteratively refines encodings through hybrid attention to a shared reference sequence. Specifically, at each timestep the encodings act as **Queries** while the reference acts as **Keys** and **Values**. We employ a hybrid attention combining self-attention and cross-attention (Vaswani et al., 2017) to compute the interaction. This extracts dependencies between words by relating them to context. Crucially, it operates on complementary LLMs-parsed and text prompt encodings, enabling informative exchange. Through iterative hybrid attention refinement, the module provides control signals capturing detailed semantic associations to guide high-fidelity generation of motion sequences conditioned on natural language description.

This fusion component aims to capture cross-modal relationships via iterative refinement of text-motion representations guided by contextual relationships extracted using hybrid attention computations. We first obtain sentence-level features $\mathbf{S}^l$ and $\mathbf{S}^t$ encoding overall semantics of the LLMs and text prompt inputs, respectively. Reference representations $\mathbf{R}^l$ and $\mathbf{R}^t$ are then generated from $\mathbf{S}^l$ and $\mathbf{S}^t$ using trainable projection matrices $\mathbf{M^l}$ and $\mathbf{M^t}$:

$$\mathbf{R}^l = \mathbf{M}^l\mathbf{S}^l, \quad \mathbf{R}^t = \mathbf{M}^t\mathbf{S}^t \tag{9}$$

We concatenate word features $\mathbf{W}^l$, $\mathbf{W}^t$ with their respective references $\mathbf{R}^t$, $\mathbf{R}^l$ and motion features $\mathbf{X}'_t$ to compute unified **Keys** and **Values** for hybrid attention. This models their mutual influence through cross-attending references:

$$\begin{aligned}\mathbf{Value} &= [\mathbf{V}^m\mathbf{X}'_t; \mathbf{V}^l[\mathbf{W}^l; \mathbf{R}^t]; \mathbf{V}^t[\mathbf{W}^t; \mathbf{R}^l]], \\ \mathbf{Key} &= [\mathbf{K}^m\mathbf{X}'_t; \mathbf{K}^l[\mathbf{W}^l; \mathbf{R}^t]; \mathbf{K}^t[\mathbf{W}^t; \mathbf{R}^l]]\end{aligned} \tag{10}$$

where $\mathbf{V}^m, \mathbf{V}^l, \mathbf{V}^t, \mathbf{K}^m, \mathbf{K}^l, \mathbf{K}^t$ denote trainable matrices. Global templates $\mathbf{G}$ extracted via softmax attention enable iterative refinement of text-motion representations:

$$\mathbf{G} = \text{softmax}(\mathbf{Key})\mathbf{Value} \tag{11}$$

Additionally, we generate a query vector at each refinement iteration to learn contextual relationships from the global template. Specifically, the motion encoding $\mathbf{X}'_t$ serves as input to produce the query vector via a trainable projection matrix $\mathbf{Q}^m$:

$$\mathbf{Query} = \mathbf{Q}^m\mathbf{X}'_t \tag{12}$$

This query vector attends to the global template $\mathbf{G}$, which encapsulates dependencies across text and motion representations inferred through iterative computations of hybrid attention over inputs from the two modalities.

$$\mathbf{Y} = \text{softmax}(\mathbf{Query})\mathbf{G} \tag{13}$$

where $\mathbf{Y} \in \mathbb{R}^{S\times D}$ gives the updated output.

By repeatedly refining encodings through extracting contextual relationships from $\mathbf{G}$, our model incrementally fuses hierarchical semantics between modalities. This enables better comprehension of textual content by grounding precise or subtle motion details in word-level semantics, enhancing performance on tasks involving understanding text through reference to implied motion concepts.

5 Experiments

5.1 Experimental Settings

Dataset. There exist some datasets for conditional generation, such as those proposed in (Plappert et al., 2016; Guo et al., 2022a, 2020; Punnakkal et al., 2021). However, datasets like (Punnakkal et al., 2021) and (Guo et al., 2020), based on action categories, cannot provide complete textual sentences, making it unsuitable for analyzing the intrinsic connections of sentence structure for our method. Instead, we use the datasets, particularly the HumanML3D dataset (Guo et al., 2022a) and the KIT-ML Motion-Language dataset (Plappert et al., 2016) for experiments. The

Methods	Publication	R Precision↑ Top 1	Top 2	Top 3	FID↓	MultiModal Dist↓	Diversity↑	MultiModality↑
TEMOS (Petrovich et al., 2022)	ECCV	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
MDM (Tevet et al., 2023)	ICLR	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799^{\pm.072}}$
MotionDiffuse (Zhang et al., 2024a)	TPAMI	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
Temporal VAE (Guo et al., 2022a)	CVPR	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MLD (Chen et al., 2023)	CVPR	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
T2M-GPT (Zhang et al., 2023a)	CVPR	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$\mathbf{9.761^{\pm.081}}$	$1.856^{\pm.011}$
MotionGPT (Jiang et al., 2024)	NeurIPS	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$0.232^{\pm.008}$	$3.096^{\pm.008}$	$9.528^{\pm.071}$	$2.008^{\pm.084}$
GraphMotion (Jin et al., 2024)	NeurIPS	$0.504^{\pm.003}$	$0.699^{\pm.002}$	$0.785^{\pm.002}$	$0.116^{\pm.007}$	$3.070^{\pm.008}$	$9.692^{\pm.067}$	$2.766^{\pm.096}$
FineMoGen (Zhang et al., 2024b)	NeurIPS	$0.504^{\pm.002}$	$0.690^{\pm.002}$	$0.784^{\pm.002}$	$0.151^{\pm.008}$	$2.998^{\pm.008}$	$9.263^{\pm.094}$	$2.696^{\pm.079}$
Att-T2M (Zhong et al., 2023)	ICCV	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$0.112^{\pm.006}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$2.452^{\pm.051}$
ReMoDiffuse (Zhang et al., 2023b)	ICCV	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$9.018^{\pm.075}$	$1.795^{\pm.043}$
Fg-T2M (Wang et al., 2023)	ICCV	$0.492^{\pm.002}$	$0.683^{\pm.003}$	$0.783^{\pm.002}$	$0.243^{\pm.019}$	$3.109^{\pm.007}$	$9.278^{\pm.072}$	$1.614^{\pm.049}$
Fg-T2M++	-	$\mathbf{0.513^{\pm.002}}$	$\mathbf{0.702^{\pm.002}}$	$\mathbf{0.801^{\pm.003}}$	$\mathbf{0.089^{\pm.004}}$	$\mathbf{2.925^{\pm.007}}$	$9.223^{\pm.114}$	$2.625^{\pm.084}$

Table 1: **Quantitative evaluation on the HumanML3D (Guo et al., 2022a) test set.** We run all the evaluation 20 times and ± indicates the 95% confidence interval. Red indicates the best result.

Methods	Publication	R Precision↑ Top 1	Top 2	Top 3	FID↓	MultiModal Dist↓	Diversity↑	MultiModality↑
TEMOS (Petrovich et al., 2022)	ECCV	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.717^{\pm.051}$	$3.417^{\pm.019}$	$10.84^{\pm.100}$	$0.532^{\pm.034}$
MDM (Tevet et al., 2023)	ICLR	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.190^{\pm.022}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$
MotionDiffuse (Zhang et al., 2024a)	TPAMI	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$11.10^{\pm.143}$	$0.730^{\pm.013}$
Temporal VAE (Guo et al., 2022a)	CVPR	$0.361^{\pm.006}$	$0.559^{\pm.007}$	$0.693^{\pm.007}$	$2.770^{\pm.109}$	$3.401^{\pm.008}$	$10.91^{\pm.119}$	$1.482^{\pm.065}$
MLD (Chen et al., 2023)	CVPR	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
T2M-GPT (Zhang et al., 2023a)	CVPR	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.92^{\pm.108}$	$1.570^{\pm.039}$
MotionGPT (Jiang et al., 2024)	NeurIPS	$0.366^{\pm.005}$	$0.558^{\pm.004}$	$0.680^{\pm.005}$	$0.510^{\pm.016}$	$3.527^{\pm.021}$	$10.35^{\pm.084}$	$2.328^{\pm.117}$
GraphMotion (Jin et al., 2024)	NeurIPS	$0.429^{\pm.007}$	$0.648^{\pm.006}$	$0.769^{\pm.006}$	$0.313^{\pm.013}$	$3.076^{\pm.022}$	$\mathbf{11.12^{\pm.135}}$	$\mathbf{3.627^{\pm.113}}$
FineMoGen (Zhang et al., 2024b)	NeurIPS	$0.432^{\pm.006}$	$0.649^{\pm.005}$	$0.772^{\pm.006}$	$0.178^{\pm.007}$	$2.869^{\pm.014}$	$10.85^{\pm.115}$	$1.877^{\pm.093}$
Att-T2M (Zhong et al., 2023)	ICCV	$0.413^{\pm.006}$	$0.632^{\pm.006}$	$0.751^{\pm.006}$	$0.870^{\pm.039}$	$3.039^{\pm.021}$	$10.96^{\pm.123}$	$2.281^{\pm.047}$
ReMoDiffuse (Zhang et al., 2023b)	ICCV	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$0.155^{\pm.006}$	$2.814^{\pm.012}$	$10.80^{\pm.105}$	$1.239^{\pm.028}$
Fg-T2M (Wang et al., 2023)	ICCV	$0.418^{\pm.005}$	$0.626^{\pm.004}$	$0.745^{\pm.004}$	$0.571^{\pm.047}$	$3.114^{\pm.015}$	$10.93^{\pm.083}$	$1.019^{\pm.029}$
Fg-T2M++	-	$\mathbf{0.442^{\pm.006}}$	$\mathbf{0.657^{\pm.005}}$	$\mathbf{0.781^{\pm.004}}$	$\mathbf{0.135^{\pm.004}}$	$\mathbf{2.696^{\pm.011}}$	$10.99^{\pm.105}$	$1.255^{\pm.078}$

Table 2: **Quantitative evaluation on the KIT-ML (Plappert et al., 2016) test set.**

HumanML3D dataset (Guo et al., 2022a) is a combination of HumanAct12 (Guo et al., 2020) and AMASS (Mahmood et al., 2019) datasets, each motion described by 3 text scripts, with an average length of about 12 words. The HumanML3D dataset (Guo et al., 2022a) includes 14616 motions and 44970 text descriptions, involving various human activities such as daily activities, sports, acrobatics, etc., with a total duration of approximately 28.59 hours. The KIT Motion-Language dataset (Plappert et al., 2016) provides a smaller-scale evaluation benchmark, with each motion sequence accompanied by one to four sentences, averaging 8 words in description length. The KIT-ML dataset (Plappert et al., 2016) consists of 3911 motion sequences and 6353 natural language descriptions, totaling approximately 10.33 hours.

Evaluation Metrics. Following the evaluation metrics (Guo et al., 2022a). (1) R-Precision (R-TOP). For each inferred text-motion pair, 31 unmatched descriptions are randomly selected from the test set. The average top-k precision is obtained by calculating and ranking the Euclidean distance between the motion and each of the 32 descriptions. (2) Frechet Inception Distance (FID). FID measures the similarity between the feature distributions extracted from the generated motions and the ground truth motions. (3) MultiModal Distance (MM-Dist). For a given description, the multimodal distance between the textual features and the corresponding generated motion features is calculated. (4) Diversity. Diversity evaluates the dissimilarity among all generated motions across all descriptions by calculating the average pairwise Euclidean distance between randomly partitioned groups of motions. (5) Multimodality. For a given text description, 32 motion sequences are randomly generated, and multimodality quantifies the dissimilarity among these generated motion sequences. We primarily focus on R-Precision and FID as key performance indicators, as they are important metrics for assessing the overall quality of generated motions.

Implementation Details. Regarding the motion encoder, we employ a 4-layer transformer with a latent dimension of 512. As for the text encoder, a frozen text encoder from CLIP ViT-B/32 is utilized, complemented by two additional transformer encoder layers. In terms of the diffusion model, the variances β_t are predefined to linearly spread from 0.0001 to 0.02, and the total number of noising steps is set at T = 1000. We use the Adam optimizer to train the model with an initial learning rate of 0.0002, gradually decreasing to 0.00002 through a cosine learning rate scheduler. The training process is conducted on 4 NVIDIA GeForce RTX 3090, with a batch size of 256 on a single GPU.

For pose representation D, we follow Guo *et al.* (Guo et al., 2022a). The pose states contain seven different parts: $(r^{va}, r^{vx}, r^{vz}, r^h, j^p, j^v, j^r)$. Here $r^{va} \in \mathbb{R}$ is the root joint's angular velocity along the Y-axis, $r^{vx}, r^{vz} \in$

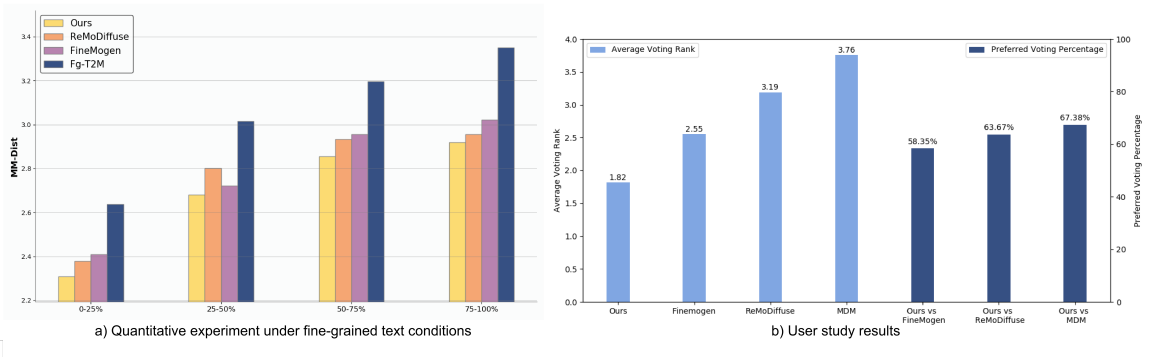

Fig. 6: **Fine-Grained Evaluation Experiment Results.** a) Evaluation of MM-Dist based on different numbers of fine-grained POSs on KIT-ML datasets (Plappert et al., 2016), where lower MM-Dist indicates better performance. The range from 0-25% to 75-100% signifies increasing difficulty levels. b) User study results on HumanML3D datasets (Guo et al., 2022a). The light blue bars on the left indicate the average voting rankings for each method, with lower rankings being better. The dark blue bars on the right represent the preference rate of Fg-T2M++ compared to other models, with higher values being better.

$\mathbb{R}$ are the root joint's linear velocities along the X-axes and Z-axes, respectively. $r^h \in \mathbb{R}$ is the height of the root height. $j^p, j^v \in \mathbb{R}^{J\times 3}$ are the positions and linear velocities of each joints. $j^r \in \mathbb{R}^{J\times 6}$ is the 6D rotation of each joint. J represents the number of joints, which are 22 in HumanML3D dataset (Guo et al., 2022a) and 21 in KIT-ML dataset (Plappert et al., 2016).

5.2 Comparison with the State of the Art

We compared our method with state-of-the-art (SOTA) models (Petrovich et al., 2022; Guo et al., 2022a; Zhang et al., 2024a; Tevet et al., 2023; Chen et al., 2023; Zhang et al., 2023a; Jiang et al., 2024; Jin et al., 2024; Zhang et al., 2024b; Zhong et al., 2023; Zhang et al., 2023b; Wang et al., 2023). Quantitative comparisons of our method with these models on the HumanML3D (Guo et al., 2022a) and KIT-ML (Plappert et al., 2016) datasets are shown in Table 1 and 2, respectively.

Compared to other methods, Fg-T2M++ achieves significantly higher scores in R-TOP, FID, and MM-Dist. These results highlight our method's proficiency in generating high-quality motion sequences that seamlessly align with the intended meanings of the provided textual prompts. Compared to SOTAs, our approach demonstrates superior performance across accuracy metrics, including R-TOP, FID, and MM-Dist. Notably, when compared to ReMoDiffuse (Zhang et al., 2023b), which employs a motion retrieval-augmented generation method aimed at matching ground truth motion distributions, our Fg-T2M++ stands out. This is attributed to our innovative approach, leveraging a sentence analysis module and LLMs parsing, which enables generated motions to better align with textual prompts. Remarkably, even without additional ground truth motion priors, Fg-T2M++ consistently outperforms competitors across all precision metrics.

Amount of POS	**Sample Prompts**
Tail 0-25%	- A person does a jumping jack. - The person runs forward fast.
Tail 25-50%	- A man jogs and stops. - The person kicked with left leg.
Tail 50-75%	- A man bends to his left several times while stretching his right arm over his head. - The person runs to their left then curves to the right and continues to run then stops.
Tail 75-100%	- A person jumps and brings both arms above his head as he spread his legs and then moves them back into the original position. - The man takes a step and picks up 3 things takes a few more steps and places one thing on the table then turns around to head back.

Table 3: Sample prompts showcasing different amount of fine-grained part-of-speech within the descriptions.

In terms of diversity metrics such as MultiModality and Diversity, the fine-grained guidance provided by our LLMs-parsed model tends to prioritize strict adherence to textual semantics. While this results in slightly weaker performance on diversity metrics, it ensures that generated motions align closely with expected textual prompts. It is important to note that prioritizing accuracy metrics strengthens the persuasiveness of our approach. After all, if generated motions fail to align with expected results, diversity metrics lose their sig-

nificance. Overall, our method demonstrates advanced experimental results and showcases the robustness of our model's performance across both datasets.

Note that when applied to general measurement methods for T2M generation, the commonly used metrics may appear moderate. This is because, when compared to real data, state-of-the-art methods achieved close scores. Hence, these metrics might not provide precise assessments, especially for more challenging complex text conditions in the generation process. In response, we specifically conducted fine-grained evaluation experiments under complex textual conditions, which we will delve into in the next section.

5.3 Fine-Grained Evaluation Experiments

We designed two evaluation experiments to assess our model's fine-grained adaptability. The first is the quantitative experiments under fine-grained text conditions. As for fine-grained texts, we rank the data according to fine-grained part-of-speech (POS) counts of "adjectives," "adverbs," "conjunctions," and "quantifiers" in sentences. We categorize all samples by ranking POS counts in ascending order, dividing the data into 0-25%, 25-50%, 50-75%, and 75-100%, from lower to higher counts. We offer two examples for each data range to better illustrate the complexity across various fine-grained parts of speech ranges, as depicted in Table 3. With higher Tail percentages, textual prompts transition from simple action sentences to complex structures containing multiple parts of speech. This evolution demands that generated motions become more fine-grained and challenging. Compared to (Zhang et al., 2023b, 2024b; Wang et al., 2023), our method outperforms the current SOTA methods, as shown in Figure 6a. Our method still maintains a better performance even though there are abundant fine-grained words in the 50-75% and 75-100% splits, indicating a better ability to capture fine-grained details.

The second part involved a user study, where we conducted comparisons with FineMoGen (Zhang et al., 2024b), ReMoDiffuse (Zhang et al., 2023b), and MDM (Tevet et al., 2023). We collected average voting ranks and user preferences to validate our earlier findings. This user study engaged 30 participants, who evaluated 15 motions generated by each method, aiming to gather comparative feedback on the question "Which method performs better in fine-grained motion modeling?". The statistical data from the user study is presented in Figure 6b. Our FG-T2M++ achieved the best voting ranking and demonstrated superior performance in preferred voting percentage compared to SOTA methods.

Methods	R Precision ↑			FID↓	MM-Dist↓
	R-TOP1	R-TOP2	R-TOP3		
Fg-T2M	$0.418^{\pm.005}$	$0.626^{\pm.004}$	$0.745^{\pm.004}$	$0.571^{\pm.047}$	$3.114^{\pm.015}$
Component Analysis of LLMs Semantic Parsing					
Only Text Prompt	$0.430^{\pm.005}$	$0.641^{\pm.003}$	$0.761^{\pm.006}$	$0.344^{\pm.021}$	$2.757^{\pm.028}$
+ Word Semantic Parsing (n1-conj)	$0.439^{\pm.010}$	$0.649^{\pm.010}$	$0.773^{\pm.014}$	$0.259^{\pm.088}$	$2.735^{\pm.013}$
+ Action Body Parsing (v1-v7)	$0.442^{\pm.006}$	$0.657^{\pm.005}$	$0.781^{\pm.004}$	$0.135^{\pm.004}$	$2.696^{\pm.011}$
Component Analysis of Hyperbolic Text Representation Module					
Standard Transformer	$0.410^{\pm.007}$	$0.611^{\pm.004}$	$0.729^{\pm.007}$	$0.724^{\pm.043}$	$3.234^{\pm.019}$
+ GCN	$0.428^{\pm.005}$	$0.641^{\pm.004}$	$0.765^{\pm.005}$	$0.357^{\pm.036}$	$2.867^{\pm.011}$
+ Hyperbolic GCN	$0.435^{\pm.006}$	$0.650^{\pm.006}$	$0.773^{\pm.005}$	$0.164^{\pm.024}$	$2.725^{\pm.014}$
+ Cross-Perception Module	$0.442^{\pm.006}$	$0.657^{\pm.005}$	$0.781^{\pm.004}$	$0.135^{\pm.004}$	$2.696^{\pm.011}$
Component Analysis of Multi-Modal Fusion Module					
Only Word-Level Feature Fusion	$0.421^{\pm.010}$	$0.635^{\pm.011}$	$0.760^{\pm.008}$	$0.281^{\pm.037}$	$2.801^{\pm.019}$
+ Word-Level Reference	$0.426^{\pm.008}$	$0.641^{\pm.006}$	$0.764^{\pm.007}$	$0.225^{\pm.026}$	$2.761^{\pm.042}$
+ Sentence-Level Feature Fusion	$0.437^{\pm.005}$	$0.651^{\pm.006}$	$0.775^{\pm.006}$	$0.167^{\pm.010}$	$2.735^{\pm.025}$
+ Sentence-Level Reference	$0.442^{\pm.006}$	$0.657^{\pm.005}$	$0.781^{\pm.004}$	$0.135^{\pm.004}$	$2.696^{\pm.011}$
Fg-T2M++	$\mathbf{0.442}^{\pm.006}$	$\mathbf{0.657}^{\pm.005}$	$\mathbf{0.781}^{\pm.004}$	$\mathbf{0.135}^{\pm.004}$	$\mathbf{2.696}^{\pm.011}$

Table 4: **Ablation of the proposed components.** All results are reported on the KIT-ML (Plappert et al., 2016) test set.

All of this highlights the adaptability of our FG-T2M++, showing its robustness in generating motions and indicating a stronger capability in capturing fine-grained details even in complex fine-grained modeling situations.

5.4 Component Analysis and Discussion

In Table 4, we conducted a comprehensive evaluation of the impact of various design components within Fg-T2M++, showcasing its performance in text-to-motion generation through extensive comparisons.

The Effectiveness of LLMs Semantic Parsing. We analyzed 15 sub-components of LLMs parsing in Table 4, with the first seven focusing on action body parsing and the remaining eight on word semantic parsing. When compared to our baseline Fg-T2M method, which utilizes only text prompts, the incorporation of LLMs analysis in Fg-T2M++ led to a significant performance enhancement. Particularly noteworthy is the greater impact of action body parsing on performance compared to word semantic parsing. Action body parsing plays a pivotal role in improving the quality of text-to-motion generation.

To further examine the role of LLMs Semantic Parsing, we conducted additional experiments under rare text conditions to validate the model's generalization performance. For rare texts, we followed the ReMoDiffuse (Zhang et al., 2023b) metric, which introduces the concept of the sample's rareness. As for a test prompt, we calculate its rareness r_p as:

$$r_p = 1 - \max_i \{< E(\text{text}_i), E(\text{prompt}) >\}, \tag{14}$$

where E represents the CLIP text encoder, text_i is the motion description in the training set, and $< \cdot, \cdot >$ denotes cosine similarity. This formula quantifies the maximum similarity between a given prompt and motion description in the training set. The higher the similarity,

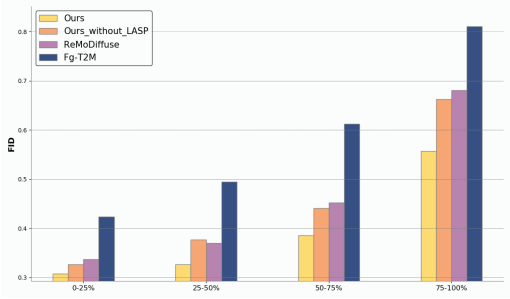

Fig. 7: **Evaluation FID based on different levels of rareness on KIT-ML (Plappert et al., 2016) datasets,** where lower FID indicates better generalization. From 0-25% to 75-100% signifies increasing difficulty levels.

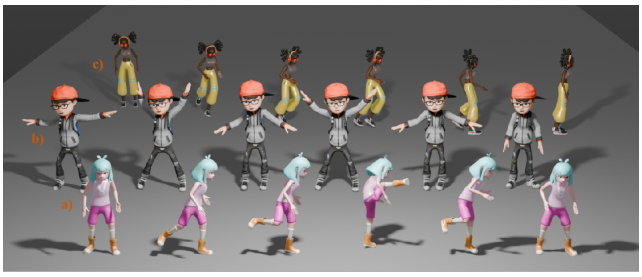

Fig. 8: **Visual results on the effects of LLMs.** Motion frames are ordered from left to right.

the lower the rareness, and vice versa. For rare texts, we rank the data based on their rareness value and divide the data into Tail 0-25%, Tail 25-50%, Tail 50-75%, and Tail 75-100%, ranging from common to rare. Compared with ReMoDiffuse (Zhang et al., 2023b) and Fg-T2M (Wang et al., 2023), in Figure 7, Fg-T2M++ generates motion sequences that more conform to the ground truth distribution, especially under rarer conditions in 75-100% splits, thus yielding significantly higher scores in FID. When our method without the LSP module, it exhibits significant degradation of FID metrics on rarer texts. This indicates that under rare text conditions, LLMs Semantic Parsing can provide more beneficial prior knowledge, thereby obtaining better generalization and generating motion that matches text more effectively.

We acknowledge that the descriptions generated by LLMs may not always be completely consistent with the GT motions, especially when the text prompts become ambiguous. However, Fg-T2M++ can still complete effective motion generation, as the fine-grained descriptions provided by current LLMs are used as reference information, not as strong constraints to limit the model.

We conducted three experiments comparing the fine-grained descriptions generated by the LLMs different from the ground truth (GT), as shown in Figure 8. The first scenario is where LLMs generated descriptions that represent the same action meaning but different variants, as shown in Figure 8a. The text prompt is: "A person kicks with the right leg". However, when LLMs parsed the detailed action of the right leg, they described it as "lifting the right leg towards a high place". The result shows that the generated figure kicks the right leg relatively high. The second scenario is where LLMs capture the overall motion semantics but do not match the specific details, as shown in Figure 8b. The text prompt is: "A person does two jumping jacks". LLMs may focus more on the action meaning brought by the jumping jacks, neglecting the depiction of "two" in the fine-grained description of each joint. However, the result shows that the figure can still complete the motion of two jumping jacks. The third scenario is where LLMs describe the motion of the wrong body parts, as shown in Figure 8c. The text prompt is: "A person runs to the left". In the current Text2Motion dataset, the directionality is mostly used to describe the left and right of the person. Therefore, this text actually describes the person running to his left side. However, due to the lack of this prior knowledge, LLMs incorrectly describe the person as running to the left side of the screen, i.e., running to the right side of the person. The result shows that the figure still completes the action of running to his left side.

The provided examples illustrate that when the fine-grained descriptions do not align with Gt as a reference, our method can robustly generate motions that are consistent with the text prompt. Nevertheless, alleviating mismatches with GT motion is still a challenge that needs to be tackled, as more precise fine-grained parsing leads to more accurate reference information. It is possible to consider adding more task priors to the LLMs' prompts to prevent some common sense issues, such as the third scenario situation mentioned above.

The Effectiveness of the Hyperbolic Text Representation Module. Our investigation explored the impact of textual syntactic structure and the utilization of hyperbolic space on text encoding in Table 4. We observed that employing a standard transformer resulted in the model struggling to capture intricate structural details within sentences, consequently leading to diminished performance. However, integrating syntactic analysis alongside graph convolutional networks, as implemented in our baseline Fg-T2M approach, significantly strengthened the model's ability to encode text, resulting in enhanced results.

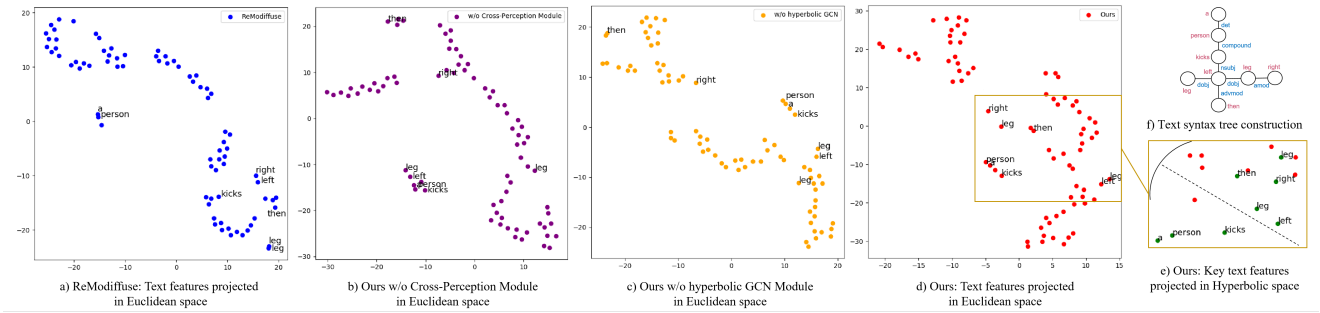

Fig. 9: **Visualization of one text sample's features in hyperbolic space and Euclidean space.** a) Text feature projection of ReMoDiffuse (Zhang et al., 2023b) into Euclidean space. b) Text feature projection of Fg-T2M++ without Cross-Perception Module into Euclidean space. c) Text feature projection of Fg-T2M++ without hyperbolic GCN Module into Euclidean space. d) Text feature projection of Fg-T2M++ into Euclidean space. e) Key text feature projection of Fg-T2M++ into hyperbolic space. f) Linguistic relationship tree structure of one text sample.

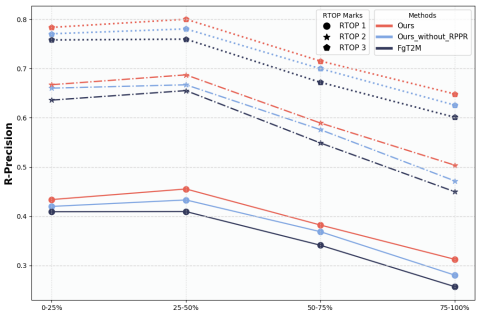

Fig. 10: **Evaluation R-TOP based on different sentence lengths on KIT-ML (Plappert et al., 2016) datasets,** where higher R-TOP indicates better performance. From 0-25% to 75-100% signifies increasing difficulty levels.

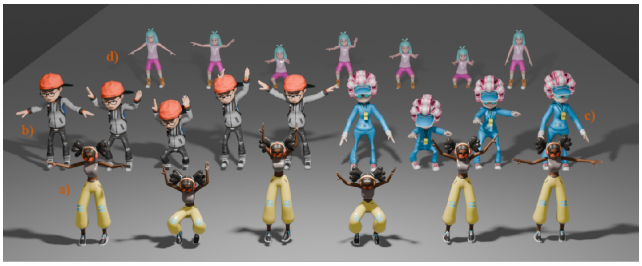

Fig. 11: **Qualitative examples on the ablation study.** Motion frames are ordered from left to right. Text prompt: A person performs two squats while lifting his arms to shoulder height and hands above his head. a): Fg-T2M++. b): Fg-T2M++ w/o multi-modal fusion module. c): Fg-T2M++ w/o hyperbolic text representation module. d): Fg-T2M++ w/o LLMs semantic parsing module.

A notable improvement in performance metrics—R-TOP, FID, and MM-Dists—was observed when hyperbolic GCN replaced the standard GCN. This highlights the efficacy of hyperbolic space in capturing tree structures and facilitating the seamless expansion of linguistic attributes throughout the generative process. Additionally, the introduction of a cross-perception module further refined the model's ability to assimilate fine textual nuances and LLMs-parsed features, leading to superior performance.

The superiority of the hyperbolic text representation module is further underscored by a visualization analysis of text features in hyperbolic space, as depicted in Figure 9. Consider the sentence "A person kicks left leg then right leg" for illustration. The linguistic features of the ReMoDiffuse (Zhang et al., 2023b) are chaotically distributed in Euclidean space, lacking a clear tree-like hierarchical organization. Additionally, features representing similar linguistic concepts, such as "left" and "right," are overly condensed, as demonstrated in Figure 9a. In stark contrast, the linguistic features of our Fg-T2M++ are arranged more logically in Euclidean space, efficiently differentiating between similar linguistic elements, for example, "left leg" and "right leg," as shown in Figure 9d. When these text language features are projected into hyperbolic space, they present a tree-like hierarchical structure, as illustrated in Figures 9e and 9f. To verify the significance of fine-grained descriptions and the function of hyperbolic GCN within the hyperbolic text representation module, we deactivate the cross-perception and hyperbolic GCN modules to study their influence on the text feature space. As depicted in Figure 9b, the absence of the

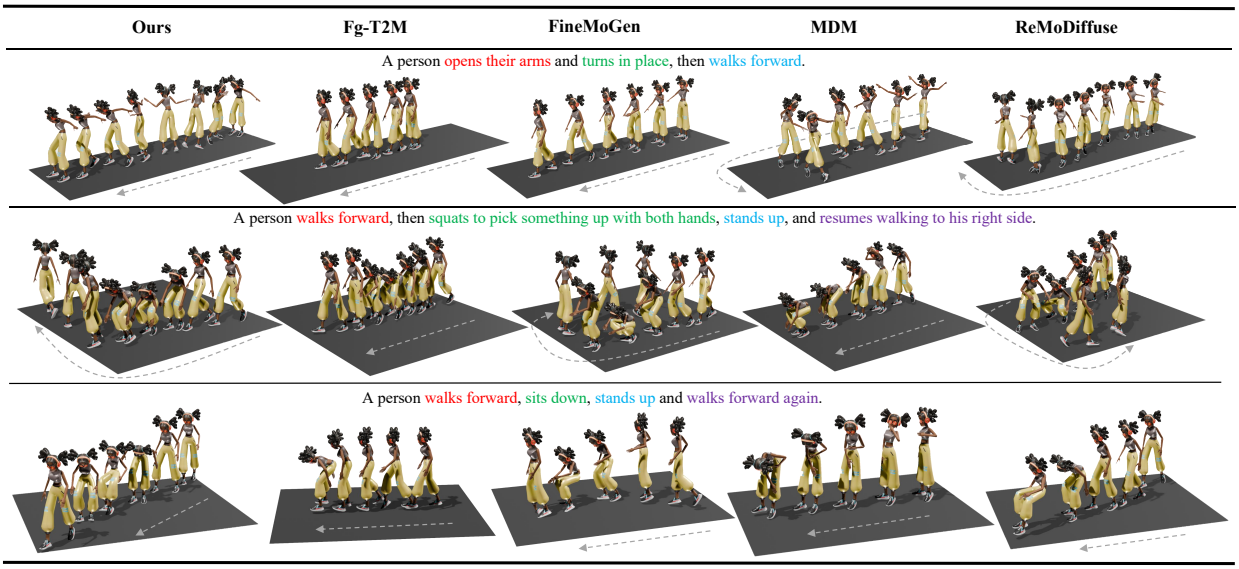

Fig. 12: **Visual results compared with existing methods.** The gray arrow represents the time axes.

cross-perception module results in a compact text feature representation that may disproportionately focus on the initial action, such as "a person kicks left leg," potentially overlooking subsequent sentence elements. Furthermore, as Figure 9c illustrates, when the hyperbolic GCN module is removed, the text features of the left leg and right leg cannot learn a clear distinction like that in Figure 9d, thus posing a great challenge to the subsequent motion generation process. This showcases that Fg-T2M++, with its Hyperbolic Text Representation Module, adeptly learns the tree-like hierarchical architecture of language, thus harnessing more effective linguistic features for enhancing motion generation.

The Effectiveness of Multi-Modal Fusion Module. Expanding upon our baseline Fg-T2M, which utilizes conventional word-level and sentence-level feature fusion methods, the integration of subtle, fine-grained insights from LLMs parsing at both the word and sentence levels substantially enriches the text-to-motion generation process by providing a deeper contextual understanding, as shown in Table 4. This strategic enhancement results in significantly improved performance. It is also observed that the fusion of word-level features, compared to sentence-level integration, has a more pronounced impact on refining and enhancing the quality of motion generation. This underscores the critical importance of linguistic analysis in advancing the fidelity of generated motions.

We further conducted additional experiments to assess the performance of the MMF Module under various lengths of text prompts. We sorted the data based on the length of the text prompts and divided it into four segments: Tail 0-25% (less than 6 words), Tail 25-50% (between 6 and 8 words), Tail 50-75% (between 8 and 10 words), and Tail 75-100% (more than 10 words), ordered from short to long. Figure 10 presents the quantitative results compared with Fg-T2M (Wang et al., 2023) and Our method without the MMF module. When the text prompts are long and complex, the performance of Fg-T2M (Wang et al., 2023), as well as our method without the MMF module, degrades significantly. However, our Fg-T2M++ shows the least degradation, demonstrating the effectiveness of our proposed MMF module, which incorporates the idea of global and local progressive fusion.

Motion Visualizations on the Ablation Study. To thoroughly assess the individual contributions of each module, we employed motion visualization for an in-depth comparative analysis, as illustrated in Figure 11. Through ablation visualization, we examined the effects of removing key components. Upon the removal of the LLM semantic parsing module, the model can still roughly complete the overall motion but lacks detail, notably failing to raise the hands above the head as specified. Excluding the multi-modal fusion module resulted in the model's limited capability to perform a single squat. The absence of the hyperbolic text representation module led to a significant decline in performance, with inaccuracies in both the number of squats and the positioning of the hands. In contrast, our full Fg-T2M++ model adeptly executed the motions as dictated by the textual descriptions, confirming the method's ability to understand complex sentences and generate high-quality motion.

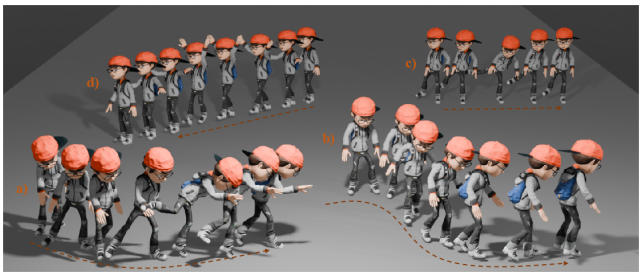

Fig. 13: **More examples of visualizations.** a): A person walks forward and lifts one leg, almost tripping over something. b): A person runs in a s-shape. c): A person raises their right leg and extends it then lowers it. d): A person is walking while raising both hands.

5.5 Qualitative Analysis

To highlight the effectiveness of Fg-T2M++, Figure 12 provides a qualitative comparison with Fg-T2M (Wang et al., 2023), FineMoGen (Zhang et al., 2024b), MDM (Tevet et al., 2023), and ReMoDiffuse (Zhang et al., 2023b). By comparison, our initial version, Fg-T2M (Wang et al., 2023), still faces challenges in capturing the intricacies within more complex sentences, often missing out on some action details. ReMoDiffuse (Zhang et al., 2023b), employing retrieval techniques, elevates the quality of motion generation and excels across action categories but encounters challenges in generating motions that precisely align with the text descriptions. MDM (Tevet et al., 2023) experiences a sharp decline in performance when faced with challenging or lengthy text prompts. FineMoGen (Zhang et al., 2024b) captures the general essence of the text but falls short of capturing finer details. Overall, our method excels in generating high-quality motions that faithfully represent the input text, surpassing these models under complex text conditions. In Figure 13, we present additional visual examples, showcasing Fg-T2M++'s robust text comprehension capabilities and its proficiency in generating intricate motions.

6 Limitations, Future Work and Conclusion

Limitations and Future Work. The effectiveness of Fg-T2M++ is closely tied to the capabilities of pre-trained large-scale language models. This reliance can present challenges, particularly in requiring applications to provide detailed and specific input formats, which to a certain extent limits its application scenarios. Additionally, the current model is limited to generating motion sequences with up to 196 frames, which restricts its application for longer sequences. Future research could focus on extending motion sequence length and achieving smooth transitions between actions to better meet real-world needs. Furthermore, modeling interactions between humans and their environments, including other people and scenes, represents another promising research direction.

Conclusion. In this paper, we introduce Fg-T2M++, a method for fine-grained text-driven human motion generation using diffusion models. Specifically, Fg-T2M++ integrates three advanced techniques: LLMs Semantic Parsing, Hyperbolic Text Representation Module, and Multi-Modal Fusion Module. By leveraging the powerful prior knowledge of LLMs to parse text prompts effectively and utilizing language relationships to construct precise language features, Fg-T2M++ achieves multi-step reasoning through hierarchical feature fusion at both global and detailed levels. Our quantitative and qualitative results demonstrate that our approach outperforms existing SOTA methods in text-driven motion generation tasks, producing high-quality, fine-grained motions that align with text prompts even under complex text conditions.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Project Number: 62272019).

Data Availability Statement. In this work, we use publicly available datasets, HumanML3D and KIT. These two datasets can be obtained at https://github.com/EricGuo5513/HumanML3D and https://drive.google.com/drive/folders/1MnixfyGfujSP-4t8w_2QvjtTVpEKr97t, respectively.

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al. (2023) Gpt-4 technical report. arXiv preprint arXiv:230308774

Ahuja C, Morency LP (2019) Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV), IEEE, pp 719–728

Athanasiou N, Petrovich M, Black MJ, Varol G (2023) Sinc: Spatial composition of 3d human motions for simultaneous action generation. arXiv preprint arXiv:230410417

Cervantes P, Sekikawa Y, Sato I, Shinoda K (2022) Implicit neural representations for variable length human motion generation. In: European Conference on Computer Vision, Springer, pp 356–372

Chen X, Jiang B, Liu W, Huang Z, Fu B, Chen T, Yu G (2023) Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 18000–18010

Desai K, Nickel M, Rajpurohit T, Johnson J, Vedantam SR (2023) Hyperbolic image-text representations. In: International Conference on Machine Learning, PMLR, pp 7694–7731

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Ghosh A, Cheema N, Oguz C, Theobalt C, Slusallek P (2021) Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1396–1406

Gilardi F, Alizadeh M, Kubli M (2023) Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:230315056

Guo C, Zuo X, Wang S, Zou S, Sun Q, Deng A, Gong M, Cheng L (2020) Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 2021–2029

Guo C, Zou S, Zuo X, Wang S, Ji W, Li X, Cheng L (2022a) Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5152–5161

Guo C, Zuo X, Wang S, Cheng L (2022b) Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision, Springer, pp 580–597

Guo C, Zuo X, Wang S, Liu X, Zou S, Gong M, Cheng L (2022c) Action2video: Generating videos of human 3d actions. International Journal of Computer Vision 130(2):285–315

Ho J, Salimans T (2022) Classifier-free diffusion guidance. arXiv preprint arXiv:220712598

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Advances in neural information processing systems 33:6840–6851

Honnibal M, Montani I (2017) spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear 7(1):411–420

Jiang B, Chen X, Liu W, Yu J, Yu G, Chen T (2024) Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems 36

Jin P, Wu Y, Fan Y, Sun Z, Yang W, Yuan L (2024) Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. Advances in Neural Information Processing Systems 36

Kalakonda SS, Maheshwari S, Sarvadevabhatla RK (2023) Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 31–36

Kao HK, Su L (2020) Temporally guided music-to-body-movement generation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 147–155

Karunratanakul K, Preechakul K, Suwajanakorn S, Tang S (2023) Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2151–2162

Kim J, Kim J, Choi S (2023) Flame: Free-form language-based motion synthesis & editing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 37, pp 8255–8263

Kochurov M, Karimov R, Kozlukov S (2020) Geoopt: Riemannian optimization in pytorch. arXiv preprint arXiv:200502819

Leng Z, Wu SC, Saleh M, Montanaro A, Yu H, Wang Y, Navab N, Liang X, Tombari F (2023) Dynamic hyperbolic attention network for fine hand-object reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14894–14904

Li R, Yang S, Ross DA, Kanazawa A (2021) Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 13401–13412

Liu Q, Nickel M, Kiela D (2019) Hyperbolic graph neural networks. Advances in neural information processing systems 32

Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2023) Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp 851–866

Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black MJ (2019) Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5442–5451

Nickel M, Kiela D (2017) Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems 30

Petrovich M, Black MJ, Varol G (2021) Action-conditioned 3d human motion synthesis with trans-

former vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10985–10995

Petrovich M, Black MJ, Varol G (2022) Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision, Springer, pp 480–497

Plappert M, Mandery C, Asfour T (2016) The kit motion-language dataset. Big data 4(4):236–252

Punnakkal AR, Chandrasekaran A, Athanasiou N, Quiros-Ramirez A, Black MJ (2021) Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 722–731

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21(1):5485–5551

Ren X, Li H, Huang Z, Chen Q (2020) Self-supervised dance video synthesis conditioned on music. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 46–54

Shafir Y, Tevet G, Kapon R, Bermano AH (2023) Human motion diffusion as a generative prior. arXiv preprint arXiv:230301418

Shen Z, Zhang M, Zhao H, Yi S, Li H (2021) Efficient attention: Attention with linear complexities. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 3531–3539

Starke S, Mason I, Komura T (2022) Deepphase: Periodic autoencoders for learning motion phase manifolds. ACM Transactions on Graphics (TOG) 41(4):1–13

Terlemez Ö, Ulbrich S, Mandery C, Do M, Vahrenkamp N, Asfour T (2014) Master motor map (mmm)—framework and toolkit for capturing, representing, and reproducing human motion on humanoid robots. In: 2014 IEEE-RAS International Conference on Humanoid Robots, IEEE, pp 894–901

Tevet G, Gordon B, Hertz A, Bermano AH, Cohen-Or D (2022) Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision, Springer, pp 358–374

Tevet G, Raab S, Gordon B, Shafir Y, Cohen-or D, Bermano AH (2023) Human motion diffusion model. In: The Eleventh International Conference on Learning Representations, URL https://openreview.net/forum?id=SJ1kSyO2jwu

Tseng J, Castellon R, Liu K (2023) Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 448–458

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in neural information processing systems 30

Wan W, Dou Z, Komura T, Wang W, Jayaraman D, Liu L (2023) Tlcontrol: Trajectory and language control for human motion synthesis. arXiv preprint arXiv:231117135

Wang Y, Leng Z, Li FW, Wu SC, Liang X (2023) Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 22035–22044

Yang M, Zhou M, Li Z, Liu J, Pan L, Xiong H, King I (2022) Hyperbolic graph neural networks: a review of methods and applications. arXiv preprint arXiv:220213852

Zhang J, Zhang Y, Cun X, Zhang Y, Zhao H, Lu H, Shen X, Shan Y (2023a) Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14730–14740

Zhang M, Guo X, Pan L, Cai Z, Hong F, Li H, Yang L, Liu Z (2023b) Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:230401116

Zhang M, Cai Z, Pan L, Hong F, Guo X, Yang L, Liu Z (2024a) Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1–15, DOI 10.1109/TPAMI.2024.3355414

Zhang M, Li H, Cai Z, Ren J, Yang L, Liu Z (2024b) Finemogen: Fine-grained spatio-temporal motion generation and editing. Advances in Neural Information Processing Systems 36

Zhong C, Hu L, Zhang Z, Xia S (2023) Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 509–519

Appendix

A User Study

Figure 14 shows the comparison page for our user study. For each text prompt, the motion is generated through different methods and randomly reshuffled. Users are required to rank their preferences for the given motions. The motions in the user study include the category of fine-grained motions, such as "a person is limping with the right leg hurt and going around in a circle" or "a person is raising his right arms above his head and then waves both hands multiple times." These text prompts contain many detailed features, requiring the generation methods to fully capture and model fine-grained features. The user study also includes the category of long sequence motions, such as "a person walks forward, then squats to pick something up with both hands, stands up, and resumes walking to his right side" or "a person walks forward, sits down, stands up, and walks forward again." These text prompts contain complex combinations of multiple actions, challenging the model's ability to learn and comprehend long sequence features from text prompts.

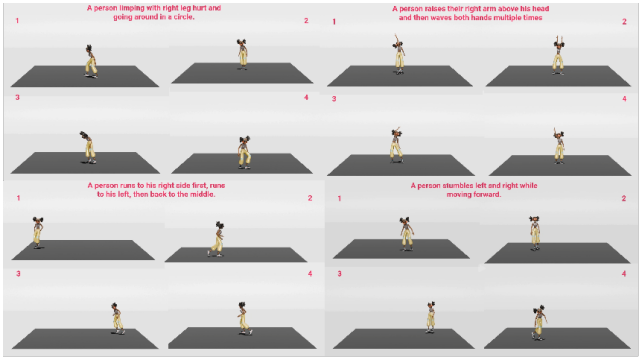

Fig. 14: Visualization comparison page of our user study.

B Failure Cases

We acknowledge some limitations in our approach, as depicted in Figure 15. While Fg-T2M++ demonstrates proficiency in capturing the subtle details embedded in text prompts, it encounters challenges when processing lengthy sentences. This occasionally leads to the omission of certain specific actions and, consequently, results in suboptimal outcomes. To address this limitation, we propose a potential solution: strategically dividing longer sentences into multiple distinct tasks for independent processing. This approach could facilitate the preservation of fine-grained characteristics in the generation of long sequence actions.

C More Analysis of Text Features

We present further visualization analysis of text features, as shown in Figure 16. The figure takes the sentence "A person sits down, stands up, and walks forward" as an example. The

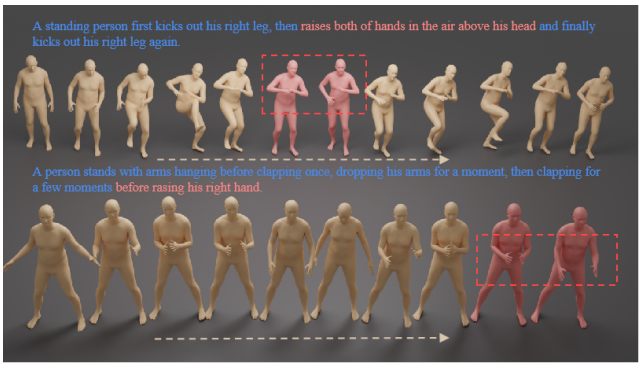

Fig. 15: Visualization of some failure cases. The arrow represents the time axes and the red box indicates the incorrect motion frames.

text features learned by ReMoDiffuse (Zhang et al., 2023b) confine a series of actions to a compact space, failing to effectively distinguish the differences between each motion. As demonstrated in the third row of Figure 12 for the ReMoDiffuse (Zhang et al., 2023b) case, it only manages to complete the action of sitting down, thereby neglecting the detailed features of the subsequent series of motions. In contrast, Fg-T2M++ learns the differences between each motion, effectively distinguishing the execution of different motions corresponding to different texts. Likewise, in the scenario depicted by text prompt 2, "a person squats to pick something up with his right hand," ReMoDiffuse (Zhang et al., 2023b) to learning textual features within a narrow text space, indicating potential overfitting during the model's training phase, which compromises its capacity for generalization. Conversely, Fg-T2M++ is adept at discerning clear and significant textual features, which in turn creates more impactful conditions that enhance the subsequent generation of motion. Therefore, it can generate high-fidelity, high-quality motion more effectively.

To verify the advantages of hyperbolic space on quantitative experiments, we calculated the geodesic distance in hyperbolic space for ReMoDiffuse (Zhang et al., 2023b) and Fg-T2M++. As shown in Table 5, D_1, D_2, and D_3 represent the average distance from the first, second, and third layer nodes to the root node in the text-tree structure, respectively. We found that the order calculated by ReMoDiffuse (Zhang et al., 2023b) is $D_2 < D_3 < D_1$, which does not conform to the hierarchical structure of the tree. In contrast, the order of Fg-t2m++ is $D_1 < D_2 < D_3$, correctly reflecting the hierarchy from the first to the third layer, revealing that our method better preserves the text-tree hierarchy.

Method	$D_1\downarrow$	$D_2\downarrow$	$D_3\downarrow$	Order
ReMoDiffuse	12.25	11.07	11.96	$D_2 < D_3 < D_1$
Fg-T2M++	**4.01**	**4.30**	**4.45**	$D_1 < D_2 < D_3$

Table 5: Comparing the performance in maintaining the hierarchical structure

Thirdly, we demonstrated the alignment between the text feature space and the motion feature space, as shown in Fig-

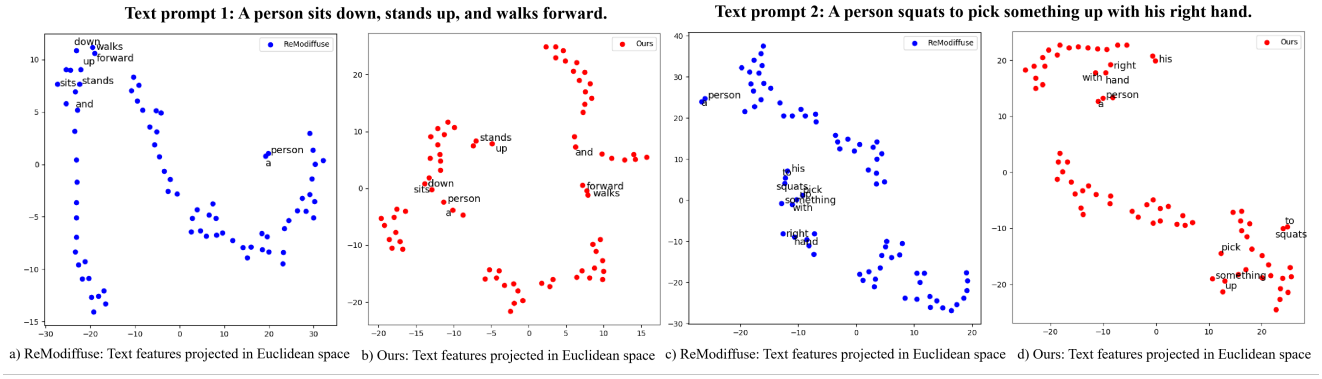

Fig. 16: **Visualization of more text samples' features in hyperbolic space and Euclidean space.** a) Text 1 feature projection of ReMoDiffuse (Zhang et al., 2023b) into Euclidean space. b) Text 1 feature projection of Fg-T2M++ into Euclidean space. c) Text 2 feature projection of ReMoDiffuse (Zhang et al., 2023b) into Euclidean space. d) Text 2 feature projection of Fg-T2M++ into Euclidean space.

ure 17. Our method achieves closer alignment between these spaces compared to ReMoDiffuse (Zhang et al., 2023b), resulting in improved cross-modal feature alignment.

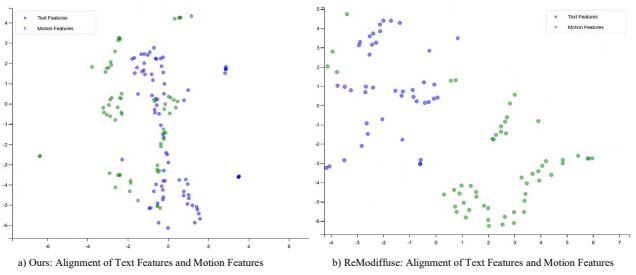

Fig. 17: The text-motion feature alignment.

Moreover, the scalability and effectiveness of hyperbolic representations over transformers have been validated in large-scale settings by works such as MERU (Desai et al., 2023), which leverage hyperbolic geometry to better preserve hierarchical relationships.

D Visual Comparison against Different Methods

We highlighted the limitations of other classes of methods when dealing with certain categories of sentences in the related work section. In this section, we provide a detailed demonstration to validate that Fg-T2M++ is capable of addressing these issues. As illustrated in Figure 18, the motions generated by our method are depicted in cool white across three images. Figure 18a demonstrates a comparison between Fg-T2M++ and the latent space alignment method, Temporal VAE (Guo et al., 2022a), revealing that Temporal VAE only produces forward-walking motion while neglecting to stumble to the left. Figure 18b shows a comparison between Fg-T2M++ and the autoregressive method TM2T (Guo et al., 2022b), which erroneously only performs a single hand swing. Figure 18c presents a comparison between Fg-T2M++ and the diffusion model method ReMoDiffuse (Zhang et al., 2023b), indicating that ReMoDiffuse fails to depict jumping and clapping simultaneously, completing only a single action. In contrast, Fg-T2M++ comprehensively generates motions consistent with the text prompts, demonstrating its superiority in handling fine-grained details.

E More Diverse examples

We present additional visualization examples, as shown in Figures 19 and 20. These demonstrate Fg-T2M++'s ability to understand complex motion descriptions and its capability to generate high-fidelity, high-quality human motion.

F Dependency Analysis of Fg-T2M++ on LLMs

We discuss the scenario where LLMs provide coarse-grained text descriptions. As shown in Figure 21a, our original text prompt is: "A person takes three steps forward." However, the coarse-grained content parsed by the LLMs lacks the number of steps. Similarly, the text prompt for Figure 21b is: "A person kicks the left leg twice," while the coarse-grained content parsed by the LLMs lacks the content of "twice." Fg-T2M++ can still generate the motion of walking and kicking, and can accurately complete the fine-grained requirement of three steps and twice. This verifies the robustness of Fg-T2M++ when the performance of LLMs is not satisfactory.

Through the analysis of the experiments, we found that the generation capability of Fg-T2M++ is not significantly affected by LLMs, i.e. when LLMs perform poorly, Fg-T2M++ does not perform poorly either. This is because Fg-T2M++ is actually more in line with the semantics of the text prompt as a whole, as the fine-grained descriptions provided by current LLMs are more used as reference information, not as strong constraints to limit the model.

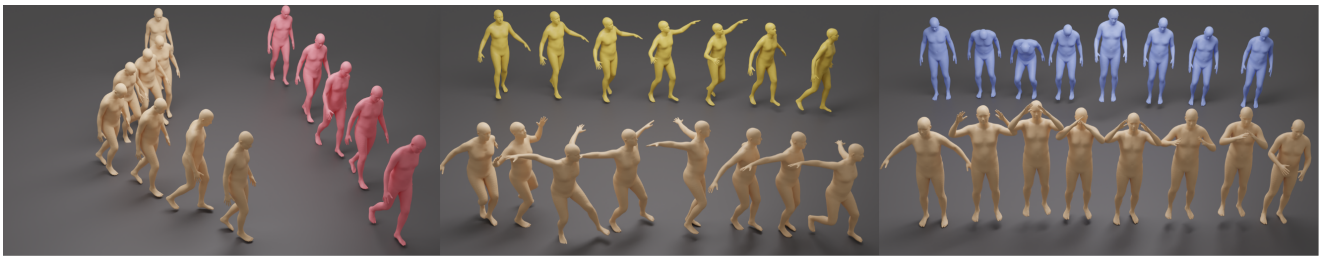

Fig. 18: **Visualization comparison with different methods.** a) Compare with latent space alignment method, Temporal VAE in red motion, under text "a person is walking forward while stumbling to the left." b) Compare with autoregressive model method, TM2T in yellow motion, under text "a person is walking with arms swinging." c) Compare with diffusion model method, ReMoDiffuse in blue motion, under text "a person is jumping while clapping."

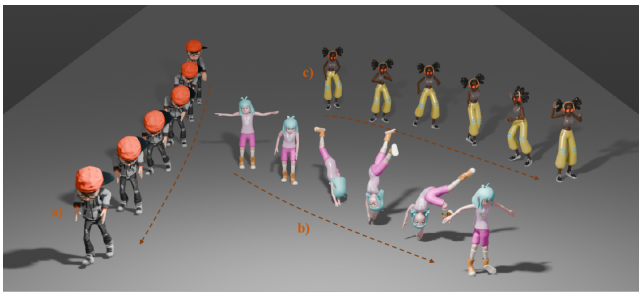

Fig. 19: More diverse qualitative samples. The arrow represents the time axes. a): A person jogs forward and looks at the ground. b): A person does a cart wheel. c): A person appears to dance.

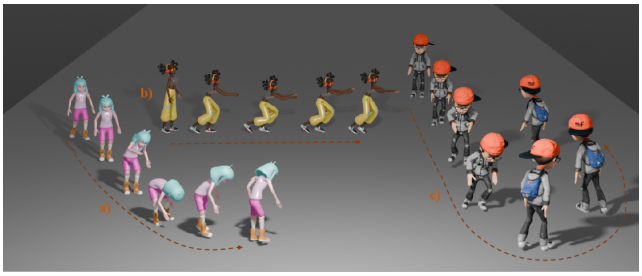

Fig. 20: More diverse qualitative samples. a): A person walks to the right and picks something up. b): A person is performing lunges. c): A person runs forward and jumps over something, then turns around.

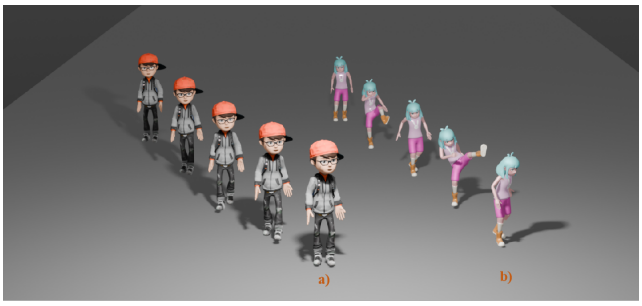

Fig. 21: Dependency analysis on LLMs

G Performance differences between GPT-3.5 and GPT-4.

To assess the performance differences between GPT-3.5 and GPT-4, we conducted two types of evaluations. First, in our quantitative evaluation, we randomly sampled 100 examples from the HumanML3D dataset (Guo et al., 2022a) and processed the text with both GPT-3.5 and GPT-4. We evaluated the motion generation quality using metrics such as R-TOP, FID, and MultiModal Dist, as shown in Table 6. The results indicated that replacing the LLMs parsing module with GPT-4 led to improvements in these metrics. This enhancement is attributed to GPT-4's ability to provide more detailed and comprehensive text analysis, which is beneficial for subsequent text feature extraction.

Method	R-TOP3 ↑	FID ↓	MultiModal Dist ↓
Ours with GPT-3.5	0.75	0.73	3.26
Ours with GPT-4	**0.77**	**0.68**	**3.15**

Table 6: The quantitative performance differences between GPT-3.5 and GPT-4.

Second, for qualitative evaluation, we visualized the motions generated from text prompts processed by both GPT-3.5 and GPT-4. As illustrated in Figure 22, we highlighted differences in parsing specific details like "left leg." GPT-3.5 failed to parse the detail of using the left foot as the root for pivot action, incorrectly using the "right foot" instead. In contrast, GPT-4's superior parsing capability allowed for a fine-grained analysis of the left foot supporting pivot action, resulting in a motion sequence that accurately matched the text description.

Fig. 22: The visualization performance differences between GPT-3.5 and GPT-4.

H LLM-parsed Fine-grained Descriptions

In this section, we present some detailed content parsed by LLMs as shown in Figure 24. These three sentences are derived from the examples in Figure 12. It can be seen that LLMs parse the original sentence into fine-grained actions for each joint part. To be more specific, take the first sentence as an example. We presented text prompts, LLM-parsed fine-grained descriptions, and corresponding motion visualizations, as illustrated in Figure 23. The parsing by LLMs of parts such as the left leg, right leg, and left arm is accurately represented in the visualizations, demonstrating enhancements in connecting text descriptions to body parts. Meanwhile, as demonstrated in the top row of Figure 12, Fg-T2M++ was the sole method to successfully achieve this precise movement. However, other approaches did not accurately reflect the arm positioning as dictated by the text description. Fg-T2M++ generates high-quality motion that conforms to the fine-grained textual description based on the LLMs-parsed output.

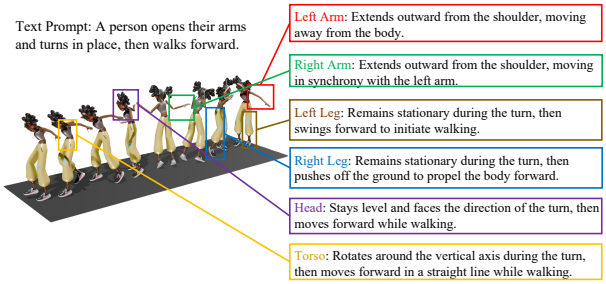

Fig. 23: Correspondence between body joints and visual motion.

Sentence-1: A person opens their arms and turns in place, then walks forward.
{"v1": "A person extends their arms outward and rotates their body on the spot, followed by a forward walking motion.",
"v2": "Left Arm: Extends outward from the shoulder, moving away from the body.",
"v3": "Right Arm: Extends outward from the shoulder, moving in synchrony with the left arm.",
"v4": "Left Leg: Remains stationary during the turn, then swings forward to initiate walking.",
"v5": "Right Leg: Remains stationary during the turn, then pushes off the ground to propel the body forward.",
"v6": "Head: Stays level and faces the direction of the turn, then moves forward while walking.",
"v7": "Torso: Rotates around the vertical axis during the turn, then moves forward in a straight line while walking.",
"n1": "",
"adj1": "forward - indicates the direction of the walk.",
"adj2": "Left Arm, Right Arm, Left Leg, Right Leg, Head, Torso",
"adv1": "in place - specifies the location of the turn, indicating no forward or backward movement.",
"adv2": "Left Arm, Right Arm, Left Leg, Right Leg, Torso",
"quant1": " ",
"quant2": " ",
"conj": "then - connects the two actions of turning in place and walking forward, indicating a sequence."}

Sentence-2: A person walks forward, then squats to pick something up with both hands, stands up, and resumes walking to his right side.
{"v1": "A person moves forward by walking, then bends the knees to squat and pick up an object with both hands, stands up straight, and continues walking in the direction of the right side.",
"v2": "Left Arm: Initially remains relatively still while walking, then lowers to pick up the object, and raises again when standing up.",
"v3": "Right Arm: Mirrors the left arm's movements, remaining still during the walk, lowering to pick up the object, and raising when standing.",
"v4": "Left Leg: Steps forward to walk, then bends at the knee to squat, and pushes up to stand.",
"v5": "Right Leg: Moves in coordination with the left leg, stepping forward, bending to squat, and extending to stand.",
"v6": "Head: Maintains a forward gaze while walking, lowers slightly when squatting, and returns to the forward position when standing.",
"v7": "Torso: Stays upright during the initial walk, leans forward when squatting, and straightens up when standing.",
"n1": "something - an unspecified object that the person picks up.",
"adj1": "forward - indicates the direction of the initial walk and the final direction of the resumed walk. both - indicates the use of both hands simultaneously",
"adj2": "forward - Left Arm, Right Arm, Left Leg, Right Leg, Head, Torso. both - Left Arm, Right Arm",
"adv1": "to pick something up - describes the purpose of the squatting action. right - indicates the direction of resumed walking",
"adv2": "Left Arm, Right Arm, Left Leg, Right Leg",
"quant1": "None",
"quant2": "",
"conj": "then - connects the actions of walking, squatting, standing, and resuming walking; and - links the actions of squatting and picking up the object; to his right side - indicates the direction of the resumed walking."}

Sentence-3: A person walks forward, sits down, stands up and walks forward again.
{"v1": "A person moves their legs to walk forward, then bends their knees and hips to sit down, rises back to a standing position using their legs, and resumes walking forward.",
"v2": "Left Arm: Swings naturally while walking, helps balance when sitting down, pushes off when standing up, swings naturally while walking again.",
"v3": "Right Arm: Swings naturally while walking, helps balance when sitting down, pushes off when standing up, swings naturally while walking again.",
"v4": "Left Leg: Alternates with the right leg in stepping forward, bends to sit, straightens to stand, alternates in stepping forward again.",
"v5": "Right Leg: Alternates with the left leg in stepping forward, bends to sit, straightens to stand, alternates in stepping forward again.",
"v6": "Head: Maintains a forward gaze, may lower slightly when sitting, and returns to the upright position when standing.",
"v7": "Torso: Stays upright while walking, leans back slightly when sitting, and straightens when standing.",
"n1": "None",
"adj1": "forward - indicates the direction of the walking motion.",
"adj2": "Left Leg, Right Leg",
"adv1": "None",
"adv2": "None",
"quant1": "None",
"quant2": "None",
"conj": "and - connects the sequence of actions: walking, sitting, standing, and walking again."}

Fig. 24: LLM-parsed fine-grained descriptions.