CoMA: Compositional Human Motion Generation with Multi-modal Agents

Shanlin Sun[*1], **Jiaqi Xu**[*3], **Gabriel de Araujo**[*1], **Shenghan Zhou**[*4], **Hanwen Zhang**[5], **Ziheng Huang**[6], **Chenyu You**[2], **Xiaohui Xie**[1]

[1]University of California, Irvine
[2]State University of New York at Stony Brook
[3]University of California, San Diego
[4]Chongqing University
[5]Huazhong University of Science and Technology
[6]Columbia University
shanlins@uci.edu, jix101@ucsd.edu, araujog@uci.edu, 20211979@stu.cqu.edu.cn, hanwenz9@usc.edu, 20216573@stu.neu.edu.cn, chenyu.you@stonybrook.edu, xhx@uci.edu

Abstract

3D human motion generation has seen substantial advancement in recent years. While state-of-the-art approaches have improved performance significantly, they still struggle with complex and detailed motions unseen in training data, largely due to the scarcity of motion datasets and the prohibitive cost of generating new training examples. To address these challenges, we introduce **CoMA**, an agent-based solution for complex human motion generation, editing, and comprehension. CoMA leverages multiple collaborative agents powered by large language and vision models, alongside a mask transformer-based motion generator featuring body part-specific encoders and codebooks for fine-grained control. Our framework enables generation of both short and long motion sequences with detailed instructions, text-guided motion editing, and self-correction for improved quality. Evaluations on the HumanML3D dataset demonstrate competitive performance against state-of-the-art methods. Additionally, we create a set of context-rich, compositional, and long text prompts, where user studies show our method significantly outperforms existing approaches.

Code — https://github.com/Siwensun/CoMA

1 Introduction

3D human motion generation has become increasingly vital across various applications, from gaming and virtual reality to robotics, spurring significant research interest. Among the emerging approaches, text-to-motion generation (Petrovich, Black, and Varol 2022; Guo et al. 2022a; Tevet et al. 2022; Zhang et al. 2023a; Guo et al. 2022c; Tevet et al. 2023; Wang et al. 2023; Karunratanakul et al. 2023b; Jiang et al. 2024; Zhang et al. 2023b, 2022) faces distinct challenges stemming from two primary factors: the inherent complexity of mapping diverse possible motions to context-rich descriptions and the limited availability of spatially and temporally complex motion data due to costly acquisition processes.

One challenge of existing methods is their performance degradation when processing context-rich motion descriptions absent from training datasets. This limitation has led to the integration of Large Language Models (LLMs) for translating general user inputs into model-compatible prompts. Notable examples include FineMoGen (Zhang et al. 2023c) and CoMo (Huang et al. 2024), which develop approaches for body part-specific instructions, while MotionGPT (Jiang et al. 2024), Motion-Agent (Wu et al. 2024) and MotionChain (Jiang et al. 2025) explore conversational interfaces for generation and editing. Building upon the use of LLMs, (Fan et al. 2025) introduces a large-scale human motion dataset and models to generate context-rich motions, and Motion-R1 (Ouyang et al. 2025) utilizes the Chain-of-Thought mechanism (Wei et al. 2022) to decompose complex prompts into structured action sequences.

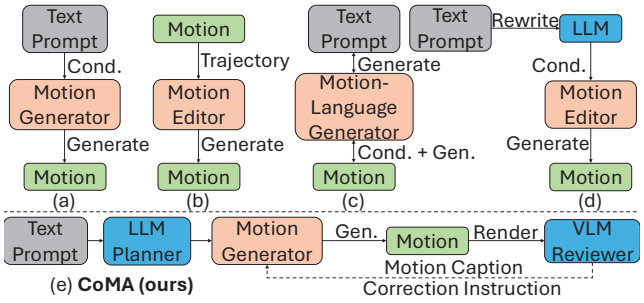

Figure 1: Illustrative architecture comparison between (a) text-conditional motion generation models, (b) keypoint- and trajectory-conditional motion editing models, (c) motion-language autoregressive models, (d) LLM-grounded motion generation models, and (e) our CoMA framework.

Despite these advances, current methods still struggle to handle spatially and temporally compositional motions, even when individual body part movements and motion segments are manageable. To generate temporally compositional motions, STMC (Petrovich et al. 2024), DiffCol-

[*]These authors contributed equally.

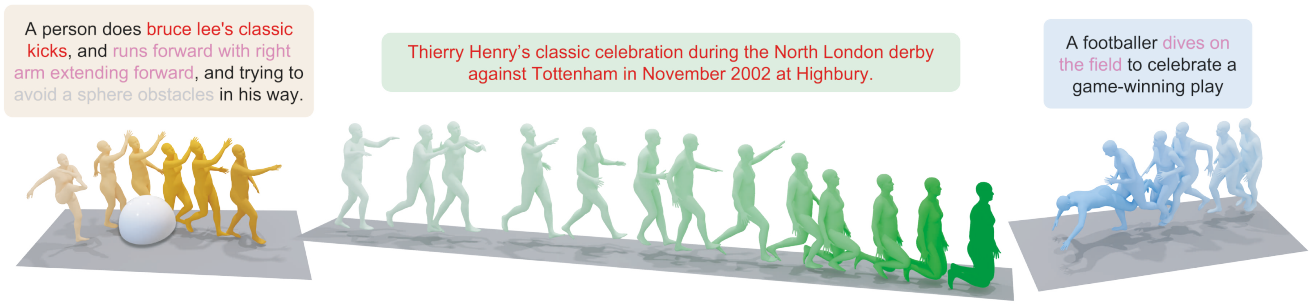

Figure 2: CoMA can generate high quality motion sequences despite challenging user expectations. *Red* indicates context-rich moves and/or poses, *Purple* indicates spatially compositional motions, and *Gray* indicates trajectory-editing instructions.

lage (Zhang et al. 2023d), and MMM (Pinyoanuntapong et al. 2024) ensure temporal coherence by stitching motion segments during denoising or predicting transition codes. For spatially compositional motions, diffusion models like FineMoGen and SINC (Athanasiou et al. 2023) leverage annotated datasets with additional body part information, while autoregressive methods like MMM, Parco (Zou et al. 2024) and Bipo (Hong et al. 2024) employ separate VQ-VAEs and transformers trained specifically to predict individual body parts.

However, none of the above methods address compositional motions while supporting context-rich motion descriptions. This motivates our proposal of CoMA, a compositional human motion generation framework with multi-model agents. As illustrated in Fig. 1, existing motion generation methods can be categorized into four main approaches: (a) text-conditional motion generation models, encompassing both diffusion-based (Tevet et al. 2023; Chen et al. 2023) and token modeling approaches (Guo et al. 2024; Pinyoanuntapong et al. 2024, 2025); (b) motion editing models that transform original motions, joint locations, or trajectories (Karunratanakul et al. 2023a,b; Xie et al. 2024; Shafir et al. 2024); (c) motion-language autoregressive models (Wu et al. 2024; Jiang et al. 2025, 2024) that integrate motion generation and understanding within unified multi-modal LLMs; and (d) LLM-grounded motion generation models (Zhang et al. 2023b,c; Huang et al. 2024) that utilize LLMs to parse user inputs into comprehensible prompts for generators.

Our work focuses on text-conditioned generation and emphasizes complex, body-part-specific generation, while also offering longer, compositional, and context-rich generations, text-based editing, and the ability to understand and correct its own generations. Tab.1 highlights our framework's distinct advantages over recent motion generation methods. Compared to state-of-the-art approaches like MoMask and MMM, CoMA excels in handling complex and unseen user inputs through LLM-based prompt re-captioning. Unlike other LLM-grounded methods such as FineMoGen and CoMo, our approach incorporates motion captioning capabilities, enabling self-correction. Moreover, in contrast to motion-language large generative models like MotionChain and Motion-Agent, CoMA automatically decomposes complex motion tasks into manageable generation and editing sub-tasks.

Our CoMA framework (Fig. 1) comprises four components: (1) **Task Planner** leverages LLM's reasoning capabilities to decompose complex motion generation tasks into manageable sub-tasks and defines comprehensive generation pipelines; (2) **Motion Generator** implements motion generation, editing, and sequence blending based on Task Planner instructions through our novel spatially-aware masked generative motion model (SPAM); (3) **Motion Reviewer** describes generated motions with our instruction-tuned video language model (MVC), evaluates motion sequence fidelity against original text prompts, and generates correction instructions through LLMs; (4) **Trajectory Editor** provides optional trajectory manipulation, generating curve functions from textual descriptions and mapping keypoints along generated trajectories to motions.

In summary, our main contributions include:

- A compositional human motion generation framework (CoMA) that handles diverse, challenging text instructions through multi-modal agent collaboration, with user studies demonstrating significant advantages over existing state-of-the-art methods.
- A spatially-aware motion generation model (SPAM) that achieves state-of-the-art performance on standard benchmarks and superior results for complex sequences.
- A motion video caption model (MVC) that demonstrates competitive performance on motion captioning tasks.

2 CoMA Overview

CoMA takes input as abstract and/or complex textual motion description and generates human motion sequences in a compositional manner; see Fig. 3. To achieve this, we design a series of collaborative multi-modal agents to decompose the process of generating human motions into simpler, singular generation tasks in different temporal segments. Each agent in CoMA is trained separately, but collaborates in a unified workflow for motion generation and editing. Detailed descriptions and pseudocode for this pipeline are provided in the Appendix.

Methods	Prompt Reasoning	Motion Caption	Composition			SC
			Spatial	Temporal	Task	
MoMask	×	×	×	×	×	×
Go to Zero	×	×	×	×	×	×
CoMo	✓	×	✓	×	×	×
Motion-R1	✓	×	×	✓	×	×
FineMoGen	✓	×	✓	✓	×	×
Mandelli et al.	✓	×	✓	✓	×	×
MotionChain	×	✓	×	✓	×	★
Motion-Agent	✓	✓	×	✓	×	★
CoMA (Ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of recent state-of-the-art methods on diverse motion-relevant tasks. *Checkmark* indicates full inclusion of the feature, *X-mark* indicates absence, and *Star* indicates incompleteness. SC stands for self-correction.

2.1 Agent Functionality

Agents in CoMA can be categorized into a high-level task planner and several low-level actors. In the following subsections, we will introduce each agent's functionality during inference. We use GPT-4o (Achiam et al. 2023) and VideoChat2 (Li et al. 2024b) for our and LLM and VLM models, respectively.

Task Planner reasons the input text prompt in three steps: text recaption, temporal segments, and task decomposition. We prompt GPT-4o to rewrite prompts to eliminate descriptions not contained in the motion datasets, which may lead to the failure in the motion generation process. GPT-4o replaces such textual abstractions and extracts hidden motion information from the original user input. As an additional safeguard against hallucination, this agent applies Retrieval Augmented Generation (Gao et al. 2024), constraining its recaptioning vocabulary to only wording found in the training dataset, for better understanding by the motion generation model. The agent may also split the rewritten text into temporally consecutive segments. This is triggered when GPT-4o identifies multiple individual motions contained in the input prompt.

Lastly, the task planner decomposes a given motion generation task into a base generation and local editing tasks. State-of-the-art motion generation models struggle with spatially compositional motions, such as "walk while raising the left hand and lowering the right hand", and even if generating such motion is possible, it remains challenging to ensure the correctness of local details. Thus, we decompose the motion generation task to generate a global motion and local body part motions separately.

Motion Generator unifies text-driven global human motion generation and local body part editing. To this end, we propose SPAM, a masked generative model where four codebooks and encoders are learned to represent four body parts, while a shared motion decoder learns to output whole human motions by fusing four local body part codes. More details can be found in Sec. 3.

Trajectory Editor is an optional agent responsible for modifying the trajectory of the motion according to textual descriptions of the trajectory. Such functionality is only triggered when explicit path control is requested. By employing Chain-of-Thought (CoT) (Wei et al. 2022) reasoning to trigger GPT-4o's spatial understanding (Li et al. 2024a), this agent generates various curve functions to produce accurate pelvis trajectories, and may also generate the required obstacles and fit such paths to avoid them. Sampling, interpolation, and resampling subsequently yield precise key points constrained in a 2D B-spline trajectory, enabling reconstruction of rotation data.

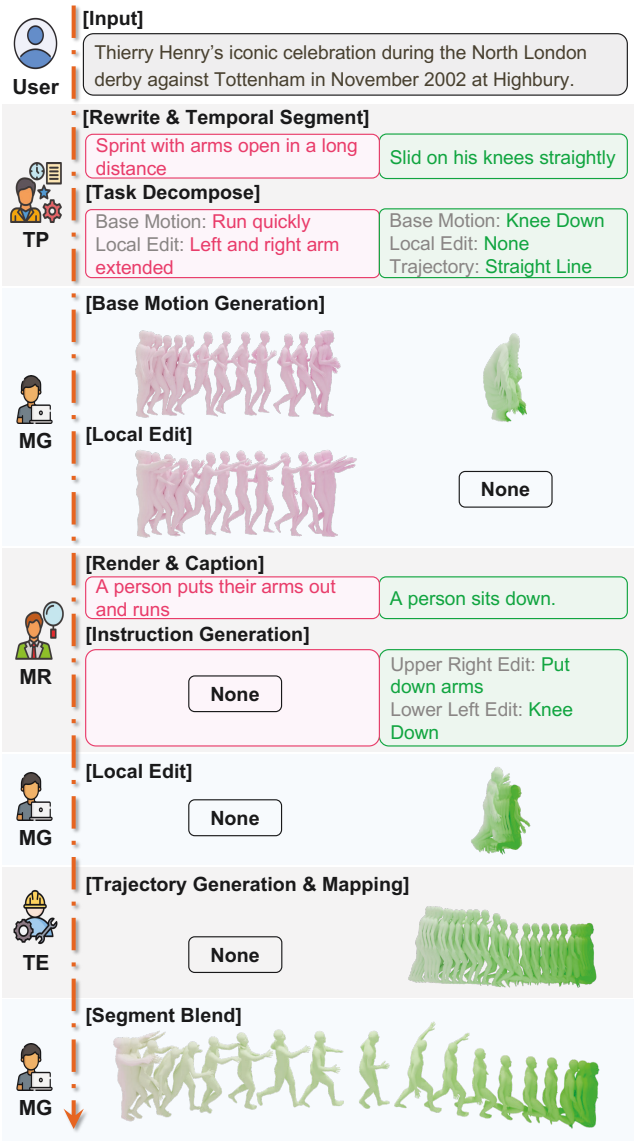

Figure 3: A real example of how our CoMA workflow generates context-rich, compositional and long motion sequence given only text prompt. TP denotes the Trajectory Planner, MG the Motion Generator, MR the Motion Reviewer and TE the Trajectory Editor.

Motion Reviewer evaluates whether a generated motion se-

quence faithfully represents the user's original text prompt. By rendering the joint outputs from SPAM and leveraging a VLM, this agent can caption motion generations and quantify their quality with respect to the initial prompt. If such output is deemed not adequate, Motion Reviewer generates specific correction instructions and returns the sequence to the Motion Generator agent for refinement.

3 Motion Generator

3.1 Preliminary Knowledge

MMM transforms motion sequences into discrete tokens using VQVAE. Given a motion sequence $\mathbf{m}_{1:N} \in \mathbb{R}^{N \times D}$, a 1D convolutional encoder $\mathcal{E}$ first encodes it into latent vectors $\mathbf{b}_{1:n} \in \mathbb{R}^{n \times d}$ with downsampling ratio n/N. Each vector is then quantized to its nearest neighbor from a codebook $\mathcal{C} = \{c_k\}_{k=1}^{K} \subset \mathbb{R}^d$ via $\mathcal{Q}(\cdot)$, producing $\tilde{\mathbf{b}}_{1:n} = \mathcal{Q}(\mathbf{b}_{1:n})$. A decoder $\mathcal{D}$ reconstructs the motion as $\tilde{\mathbf{m}} = \mathcal{D}(\tilde{\mathbf{b}})$, with the codebook indices serving as discrete motion tokens. MoMask extends this using residual vector quantization (RVQ) (Zeghidour et al. 2021) to produce multiple token layers. Starting with $r_0 = \mathbf{b}$, each layer v computes:

$$\tilde{\mathbf{b}}^v = \mathrm{Q}(\mathbf{r}^v), \quad \mathbf{r}^{v+1} = \mathbf{r}^v - \tilde{\mathbf{b}}^v \tag{1}$$

where $v = 0, \ldots, V$. The final latent approximation $\sum_{v=0}^{V} \tilde{\mathbf{b}}^v$ is then decoded through $\mathcal{D}$.

For text-guided generation, MoMask uses two transformers: a masked transformer generating base-layer tokens $t^0_{1:n}$ with masking schedule $\gamma(\tau) = \cos(\frac{\pi\tau}{2})$, and a residual transformer sequentially predicting tokens for layers 1 to V. Both employ classifier-free guidance during inference, computing logits as $\omega_g = (1+s)\cdot\omega_c - s\cdot\omega_u$, where ω_c and ω_u are conditional and unconditional predictions.

3.2 SPAM

CoMA aims to deliver a unified Motion Generator agent that not only generates complex human motions from global text prompts in one shot, but also understands granular editing instructions to modify specified body parts. We propose a Spatially-Aware Masked Generative Motion Model (SPAM), which processes both local and global text prompts to coherently generate and edit four body parts (right/left upper/lower). Next, We focus on explaining the key differences between our model and the original MoMask architecture.

Spatially-Aware Residual VQVAE Our spatially-aware VQVAE consists of four encoders, four separate quantizers and one shared decoder, as is shown in Fig. 4(a). Each body part has its own encoder and codebook, which converts the corresponding motion sequence $\mathbf{m}^i \in \mathbb{R}^{N \times D_i}$ into a latent vector sequence $\mathbf{b}_i \in \mathbb{R}^{n \times d}$. Thus, one motion sequence can be represented by four tuples of body parts motion tokens:

$$\mathbf{B} = [\mathbf{b}_i]_{i=1}^{4} \in \mathbb{R}^{4 \times n \times d} \tag{2}$$

each of which is generated by a corresponding quantizer and encoder. Finally, the four body parts motion tokens are concatenated, which will be decoded into motion space by a shared decoder, generating whole body motions:

$$\tilde{\mathbf{m}} = \hat{\mathcal{D}}\left(\text{concat}\left(\left[\mathcal{Q}_i\left(\mathcal{E}_i\left(\mathbf{m}^i\right)\right)\right]_{i=1}^{4}\right)\right) \tag{3}$$

Following MoMask, we train the residual motion VQ-VAEs via a motion reconstruction loss combined with a latent embedding loss at each quantization layer:

$$\mathcal{L}_{\text{rvq}} = \|\mathbf{m} - \tilde{\mathbf{m}}\|_1 + \beta \sum_{v=1}^{V} \|\mathbf{R}^v - \text{sg}[\mathbf{B}^v]\|_2^2 \tag{4}$$

where $\mathbf{R}^v = [r_i^v]_{i=1}^4$ denotes the residual tokens tuples, $\text{sg}[\cdot]$ denotes the stop-gradient operation, and β is a weighting factor for the embedding constraint. This framework is optimized with a straight-through gradient estimator (Van Den Oord, Vinyals et al. 2017), and our codebooks are updated via exponential moving average and codebook reset following T2M-GPT (Zhang et al. 2023a).

Spatially-Aware Motion Transformer Our Spatially-Aware Transformer models both base-layer motion token tuples $T^0 = [t_i^0]_{i=1}^4 \in \mathbb{R}^{4 \times n}$ and residual-layer motion token tuples $[T^v]_{v=1}^V \in \mathbb{R}^{V \times 4 \times n}$ using base transformer f_θ and residual transformer f_ϕ, where each base token tuple t_i^0 represents a distinct body part. Inspired by video classification architectures (Bertasius, Wang, and Torresani 2021), we implement a factorized space-time self-attention mechanism for motion generation. As is shown in Fig. 4(b), our SPAM transformer splits the attention computation into two sequential steps: spatial attention across body parts, followed by temporal attention across time steps. In this design, tokens first attend to others within the same time step through spatial self-attention, capturing inter-part relationships. Subsequently, temporal self-attention is applied to each spatial position across time steps to model temporal dependencies. This factorized approach reduces computational complexity while maintaining expressiveness by separately modeling spatial and temporal relationships.

Base Transformer As shown in the Fig. 4(b), given masked motion tuples $\hat{T}^0$ and text prompts $P = [\mathbf{p}^i]_{i=1}^4$, where $\mathbf{p}^i$ describes body part i, the base transformer predicts the masked tokens. The text prompts can be either identical global descriptions or distinct part-specific instructions. We extract text features using CLIP (Tevet et al. 2022). The base transformer f_θ is trained to minimize:

$$\mathcal{L}_{\text{base}} = \sum_{\hat{T}_k = [\text{MASK}]} -\log f_\theta(T_k^0 \mid \hat{T}^0, P) \tag{5}$$

Residual Transformer mirrors the base transformer's architecture but maintains V separate embedding layers. Given a randomly selected layer $j \in [1, V]$, it embeds and sums tokens from preceding layers $T^{0:j-1}$, then predicts tokens for layer j conditioned on these embeddings, text P, and layer index j. The residual transformer f_ϕ is trained to minimize:

$$\mathcal{L}_{\text{res}} = \sum_{j=1}^{V} \sum_{i=1}^{n} -\log f_\phi(T_i^j \mid T_i^{0:j-1}, P, j) \tag{6}$$

Motion Editing For complex motions, SPAM may struggle to generate satisfactory results in one shot, requiring a combination of simple generation and fine editing, which our CoMA system supports collaboratively. SPAM supports

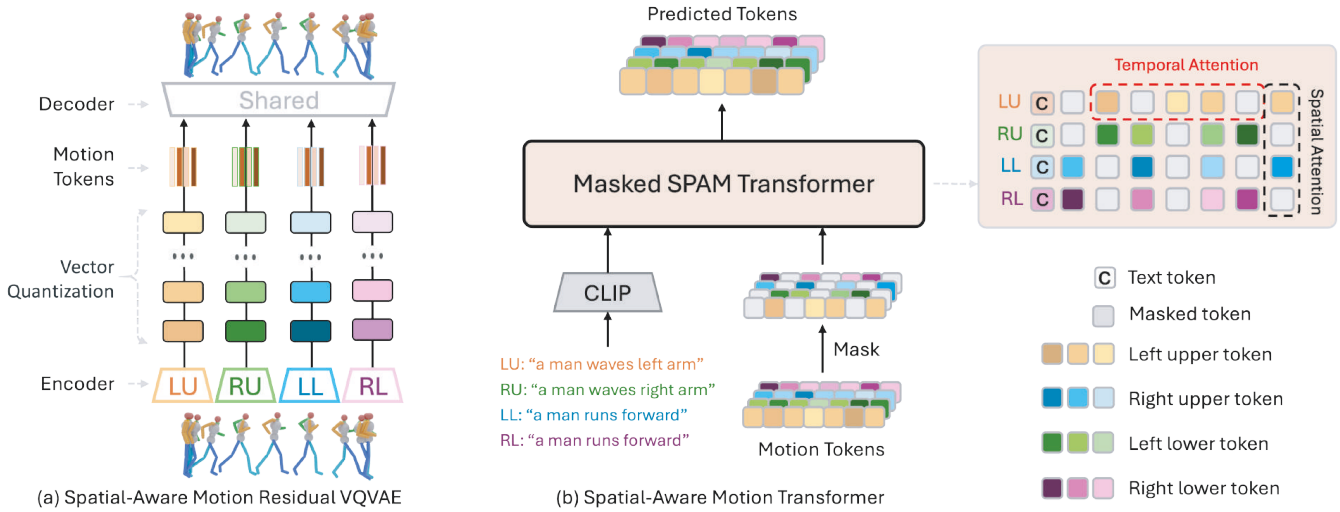

Figure 4: SPAM overview. (a) Motion sequence is decomposed into four body parts: left upper (LU), right upper (RU), left lower (LL), and right lower (RL). Each part is tokenized through separate RVQs and reconstructed into a whole-body motion through a shared decoder. (b) Base-layer motion tokens are randomly masked, while local/global text prompts are encoded separately and concatenated with corresponding motion tokens. The Masked SPAM Transformer is trained to predict the masked tokens. The residual transformer follows a similar architecture and is omitted for brevity.

multiple motion editing tasks to iteratively refine the motion. All of the editing tasks below do not require additional training and can seamlessly integrate with each other.

In-between Editing After the user sets frames $\alpha : \beta$ to be edited, corresponding token tuples $\mathbf{T}^0_{\alpha:\beta}$ will be replaced with [MASK]. Our SPAM will fill in these [MASK] tokens and generate a natural animation. Motion Reviewer agent can automatically select the keyframes to be edit.

Body Part Editing Our model supports text-driven editing of four body parts: left upper, right upper, left lower, and right lower. After specifying the body parts J to be edited, the corresponding token sequences $[\mathbf{t}^0_j]_{j\in J}$ will be replaced with `[MASK]`. To ensure a natural connection between the other body parts and the edited parts, we introduce random `[MASK]` tokens into the other body parts. The Motion Reviewer agent will automatically select the parts to be edited until the desired result is achieved.

Blend Editing Inspired by MMM, given two sequences of motions, the model will generate transition motion tokens conditioned on the end of the first sequence and the start of the second one.

4 Experiments

We evaluated CoMA from two perspectives: quantitative performance on the standard HumanML3D benchmark (Guo et al. 2022b), and qualitative assessment through human studies focused on complex motion generation.

4.1 Text-to-Motion Generation

Setup HumanML3D contains 14,616 motions extracted from multiple source datasets, with each motion paired with 3 textual descriptions, totaling 44,970 possible prompts. We use the standard split comprising 23,384 training samples, 1,460 validation samples, and 4,384 test samples.

Following prior works in motion generation (Tevet et al. 2023; Guo et al. 2024), we adopt standard evaluation metrics: Frechet Inception Distance (FID) for measuring distributional similarity between generated and ground truth motions, R-Precision and Multimodal Distance (matching score) for assessing text-motion semantic alignment, and Multimodality for quantifying generation diversity.

We compare our method against state-of-the-art approaches across three categories: diffusion-based methods (FineMoGen, CoMo and MDM (Tevet et al. 2023)), masked generation methods (MMM and MoMask), and autoregressive methods (T2M-GPT and MotionGPT), as well as large motion-language models (MotionChain and Motion-Agent).

Generation Results Tab. 2 presents results on the standard HumanML3D benchmark. Our SPAM achieves top-3 performance in FID, Multimodal Distance, and R-Precision metrics. Notably, we achieve the best performance in Top-1 and Top-2 R-precision while ranking second in Top-3, demonstrating our model's superior instruction-following capability. Additionally, we demonstrate that SPAM is better suited for fine-grained text, as detailed in Appendix.

4.2 Text-guided Motion Editing

SPAM's spatial understanding enables precise motion editing across four main body parts. While existing methods like MMM support basic upper/lower body division, they struggle with fine-grained editing tasks. Fig. 5 demonstrates this limitation: given the input motion "A person waves his left hand in greeting," when editing to include "while rasing his right hand to his head," MMM fails to preserve the original

Methods	R Precision↑			FID↓	MultiModal Dist↓	MultiModality↑
	Top 1	Top 2	Top 3			
MDM (Tevet et al. 2023)	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$2.799^{\pm.072}$
T2M-GPT (Zhang et al. 2023a)	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$1.831^{\pm.048}$
CoMo (Huang et al. 2024)	$0.502^{\pm.002}$	$0.692^{\pm.007}$	$0.790^{\pm.002}$	$0.262^{\pm.004}$	$3.032^{\pm.015}$	$1.013^{\pm.046}$
FineMoGen (Zhang et al. 2023c)	$0.504^{\pm.002}$	$0.690^{\pm.002}$	$0.784^{\pm.002}$	$0.151^{\pm.008}$	$2.998^{\pm.008}$	$2.696^{\pm.079}$
MotionChain (Jiang et al. 2025)	$0.504^{\pm.003}$	$0.695^{\pm.003}$	$0.790^{\pm.003}$	$0.248^{\pm.009}$	$3.033^{\pm.010}$	$1.715^{\pm.066}$
Motion-Agent (Wu et al. 2024)	$0.515^{\pm.004}$	-	$0.801^{\pm.004}$	$0.230^{\pm.009}$	$2.967^{\pm.020}$	-
MotionGPT (Jiang et al. 2024)	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$0.232^{\pm.008}$	$3.096^{\pm.008}$	$2.008^{\pm.084}$
MMM (Pinyoanuntapong et al. 2024)	$0.515^{\pm.002}$	$0.708^{\pm.002}$	$0.804^{\pm.002}$	$0.089^{\pm.005}$	$2.926^{\pm.007}$	$1.226^{\pm.035}$
MoMask (Guo et al. 2024)	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	$1.241^{\pm.040}$
Motion-R1 (Ouyang et al. 2025)	$0.515^{\pm.003}$	$0.719^{\pm.002}$	$0.818^{\pm.002}$	$0.201^{\pm.044}$	$2.854^{\pm.010}$	$2.317^{\pm.105}$
SPAM	$0.526^{\pm.003}$	$0.713^{\pm.003}$	$0.805^{\pm.003}$	$0.092^{\pm.006}$	$2.939^{\pm.008}$	$0.924^{\pm.039}$

Table 2: Quantitative evaluation on the HumanML3D test set. ± indicates a 95% confidence interval.

left-hand motion, while Motion-Agent struggles to perform two motions simultaneously. In contrast, CoMA not only allows specific body part editing but also provides automatic modification suggestions through VLM integration.

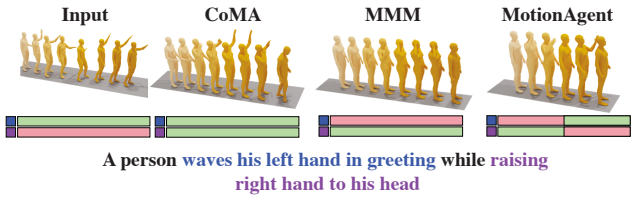

Figure 5: Editing abilities of CoMA, MMM and Motion-Agent with input motion 'A person waves his left hand in greeting'.

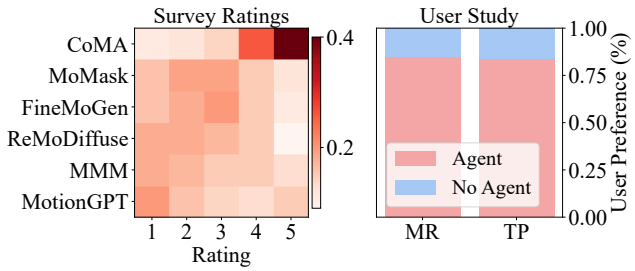

Figure 6: User Study Results. Heatmap rating for each video is in the 1-5 range, where higher is better. The agent ablation plot shows user preference for sequences generated with and without each listed agent. MR stands fr Motion Reviewer, and TP for Task Planner.

4.3 Motion Caption Results

Following other state-of-the-art methods (Guo et al. 2022c; Jiang et al. 2024), we evaluate our motion captioning performance using standard NLP metrics: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and

Model	Bleu@1 ↑	Bleu@4 ↑	Rouge ↑	Cider ↑	Bert Score ↑
TM2T	48.90	8.27	38.1	15.80	32.2
MotionGPT	48.20	12.47	37.4	29.20	32.4
MotionChain	48.10	12.56	33.9	33.70	36.9
MotionLLM	54.53	17.65	48.7	33.74	42.6
MVC (ours)	60.05	20.98	45.79	44.03	40.12

Table 3: Quantitative comparison of motion captioning on HumanML3D.

BERTScore (Zhang et al. 2019). For fair comparison, we adopt Motion-Agent's evaluation approach, using unprocessed ground truth text that ignores tense and plural variations. As demonstrated in Tab. 3, our MVC model, with implementation details in Appendix, achieves superior performance in Bleu and Cider metrics, indicating its ability to generate precise and accurate motion descriptions.

4.4 User Study

Setup To evaluate performance on complex motions, we conducted a user study comparing whole-pipeline CoMA against state-of-the-art open-sourced approaches: MoMask, ReMoDiffuse (Zhang et al. 2023b), MMM, MotionGPT and FineMoGen. We designed 42 challenging prompts featuring long, context-rich, spatially compositional motion descriptions (detailed in Appendix). The study involved 96 participants evaluating motion sequences across multiple test cases, scoring both motion quality and text-prompt alignment. We also introduce a novel Motion Alignment Score (MAS) metric, which measures video-text embedding similarity using InternVideo2 (Wang et al. 2024). This metric enables evaluation of any motion with minimal samples by comparing embeddings from the video encoder (for rendered motion) and text encoder (for prompts).

Results Fig. 6 presents evaluation results across average score, ranking, and individual agent ablation. CoMA consistently outperforms existing approaches across all criteria. In direct comparison with MoMask, our method shows superior capability in complex motion generation. MAS scores

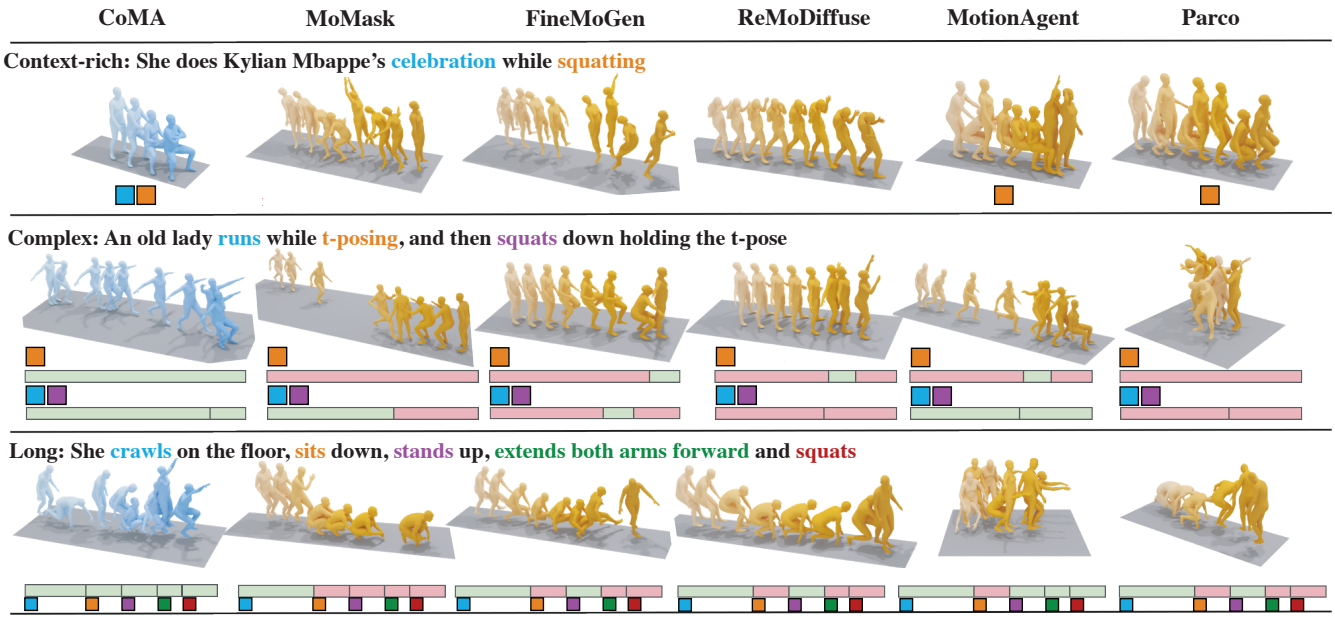

Figure 7: A qualitative comparison between CoMA and state-of-the-art models on three challenging tasks: context-rich motion generation, complex motion generation, and long motion generation.

align with the reported ranking of user preference. Visual comparisons in Fig. 7 demonstrate our method's significant advantages in generating context-rich, complex, and extended motion sequences. As shown in the context-rich example, while Motion-Agent (Wu et al. 2024) and Parco (Zou et al. 2024) generate squat motions, the figure fails to perform Mbappe's signature celebration. For complex motions, only CoMA maintains a T-pose throughout and performs the correct motion. For long motions, CoMA completes the entire sequence clearly, whereas Motion-Agent(Wu et al. 2024) only partially executes them in the wrong order.

4.5 Ablation Study

We ablate the importance of each agent using the MTT (Petrovich et al. 2024) complex motion dataset and qualitative results. Following the STMC (Petrovich et al. 2024) evaluation protocols, Tab. 4 demonstrate per-agent performance improvements. We compare two state-of-the-art models on MTT: STMC's MotionDiffuse trained on HumanML3D, and EnergyMoGen (Zhang, Fan, and Yang 2025), which leverages energy-based models for semantic composition.

In addition, we conducted a separate study for individual agent ablation with a cohort of 18 participants, results shown in Fig. 6.

Metric	GT	M1	M2	M3	Task Planner		+Reviewer
					+Temporal	+Task	
FID ↓	0.00	0.53	0.58	0.57	0.48	0.49	0.50
R@1 ↑	55.0	24.8	14.0	12.0	30.1	32.0	33.1
R@3 ↑	73.3	46.7	26.3	24.2	49.6	53.1	54.6
M2T ↑	0.74	0.66	0.57	0.57	0.66	0.67	0.67
M2M ↑	1.00	0.63	0.56	0.57	0.63	0.63	0.63

Table 4: Agent ablation Study on MTT dataset. Task planner and motion reviewer improve text-motion alignment. Models: M1 is STMC, M2 is E-MOGEN and M3 is SPAM.

5 Discussion

We proposed CoMA, a multi-modal compositional human motion generation framework that refines complex human motion generations from textual descriptions. With four multi-modal agents powered by an LLM, VLM, and a spatially-aware generative motion model, CoMA enables longer generations, text-driven editing, motion composition, and self-correction, consistently delivering higher-quality results on standard benchmarks and challenging cases.

Limitations and future work. Our framework leverages the reasoning capability of LLMs to generate fine-grained motion sequences. Although our empirical results are promising, relying on LLMs may lead to hallucinations. Integrating Chain-of-Thought (CoT) and Retrieval-Augmented Generation (RAG) mitigates this issue, but we foresee room for further exploration. Future work could introduce task-specific fine-tuning of the task planner agent and learn generalized motion generation policies via scalable reinforcement learning (Xu et al. 2024; Guo et al. 2025). Distilling larger models' reasoning abilities into smaller task planner agents would improve inference efficiency. Additionally, CoMA can serve as a reliable data engine to generate complex and diverse text-motion paired data for training unified end-to-end multi-modal language models that support motion generation, understanding, and editing.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Athanasiou, N.; Petrovich, M.; Black, M. J.; and Varol, G. 2023. SINC: Spatial Composition of 3D Human Motions for Simultaneous Action Generation. In *ICCV*.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Fan, K.; Lu, S.; Dai, M.; Yu, R.; Xiao, L.; Dou, Z.; Dong, J.; Ma, L.; and Wang, J. 2025. Go to Zero: Towards Zero-shot Motion Generation with Million-scale Data. arXiv:2507.07095.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1910.

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

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022b. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022c. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *ECCV*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hong, S.-E.; Lim, S.; Hwang, J.; Chang, M.; and Kang, H. 2024. BiPO: Bidirectional Partial Occlusion Network for Text-to-Motion Synthesis. *arXiv preprint arXiv:2412.00112*.

Huang, Y.; Wan, W.; Yang, Y.; Callison-Burch, C.; Yatskar, M.; and Liu, L. 2024. CoMo: Controllable Motion Generation through Language Guided Pose Code Editing. arXiv:2403.13900.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2024. MotionGPT: Human Motion as a Foreign Language. *Advances in Neural Information Processing Systems*, 36.

Jiang, B.; Chen, X.; Zhang, C.; Yin, F.; Li, Z.; Yu, G.; and Fan, J. 2025. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, 54–74. Springer.

Karunratanakul, K.; Preechakul, K.; Aksan, E.; Beeler, T.; Suwajanakorn, S.; and Tang, S. 2023a. Optimizing Diffusion Noise Can Serve As Universal Motion Priors. In *arxiv:2312.11994*.

Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023b. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2151–2162.

Li, F.; et al. 2024a. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *AAAI*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Ouyang, R.; Li, H.; Zhang, Z.; Wang, X.; Zhu, Z.; Huang, G.; and Wang, X. 2025. Motion-R1: Chain-of-Thought Reasoning and Reinforcement Learning for Human Motion Generation. arXiv:2506.10353.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 480–497. Springer.

Petrovich, M.; Litany, O.; Iqbal, U.; Black, M. J.; Varol, G.; Bin Peng, X.; and Rempe, D. 2024. Multi-track timeline control for text-driven 3d human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1911–1921.

Pinyoanuntapong, E.; Saleem, M. U.; Wang, P.; Lee, M.; Das, S.; and Chen, C. 2025. BAMM: bidirectional autoregressive motion model. In *European Conference on Computer Vision*, 172–190. Springer.

Pinyoanuntapong, E.; Wang, P.; Lee, M.; and Chen, C. 2024. MMM: Generative Masked Motion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022. MotionCLIP: Exposing Human Motion Generation to CLIP Space. *arXiv preprint arXiv:2203.08063*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, Y.; Leng, Z.; Li, F. W.; Wu, S.-C.; and 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*, 22035–22044.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Chen, G.; Pei, B.; Zheng, R.; Xu, J.; Wang, Z.; et al. 2024. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Q.; Zhao, Y.; Wang, Y.; Liu, X.; Tai, Y.-W.; and Tang, C.-K. 2024. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. arXiv:2405.17013.

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

Xu, C.; Li, Q.; Luo, J.; and Levine, S. 2024. RLDG: Robotic Generalist Policy Distillation via Reinforcement Learning. *arXiv preprint arXiv:2412.09858*.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2021. Soundstream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.

Zhang, J.; Fan, H.; and Yang, Y. 2025. EnergyMoGen: Compositional Human Motion Generation with Energy-Based Diffusion Model in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17592–17602.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023a. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001*.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023b. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. *arXiv preprint arXiv:2304.01116*.

Zhang, M.; Li, H.; Cai, Z.; Ren, J.; Yang, L.; and Liu, Z. 2023c. FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing. *NeurIPS*.

Zhang, Q.; Song, J.; Huang, X.; Chen, Y.; and yu Liu, M. 2023d. DiffCollage: Parallel Generation of Large Content with Diffusion Models. In *CVPR*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zou, Q.; Yuan, S.; Du, S.; Wang, Y.; Liu, C.; Xu, Y.; Chen, J.; and Ji, X. 2024. ParCo: Part-Coordinating Text-to-Motion Synthesis. *arXiv preprint arXiv:2403.18512*.