MotionChain: Conversational Motion Controllers via Multimodal Prompts

Biao Jiang[1,2⋆], Xin Chen[2⋆⋆], Chi Zhang[2], Fukun Yin[1], Zhuoyuan Li[1], Gang Yu[2], and Jiayuan Fan[1⋆⋆⋆]

[1] Fudan University
[2] Tencent

Abstract. Recent advancements in language models have demonstrated their adeptness in conducting multi-turn dialogues and retaining conversational context. However, this proficiency remains largely unexplored in other multimodal generative models, particularly in human motion models. By integrating multi-turn conversations in controlling continuous virtual human movements, generative human motion models can achieve an intuitive and step-by-step process of human task execution for humanoid robotics, game agents, or other embodied systems. In this work, we present MotionChain, a conversational human motion controller that generates continuous and long-term human motion through multimodal prompts. Specifically, MotionChain consists of multi-modal tokenizers that transform various data types such as text, image, and motion, into discrete tokens, coupled with a Vision-Motion-aware Language model. By leveraging large-scale language, vision-language, and vision-motion data to assist motion-related generation tasks, MotionChain thus comprehends each instruction in multi-turn conversation and generates human motions followed by these prompts. Extensive experiments validate the efficacy of MotionChain, demonstrating state-of-the-art performance in conversational motion generation, as well as more intuitive manners of controlling and interacting with virtual humans.

Keywords: 3D Motion · Motion Generation · Text-to-Motion

1 Introduction

The success of large language models (LLMs) [60, 62, 93, 94, 122] has sparked significant interest in the development of multi-modal language models. These models aim to transfer instruction-following and zero-shot abilities to other modalities tasks, such as image-language models [1, 50, 107, 124], video-language models [1, 43, 44, 113], and 3D-language models [12, 26, 28, 108]. However, a comprehensive model that can perceive visual input and generate continuous motion through multi-turn conversations has not yet been developed. Such a

⋆ Work done while Biao Jiang was a Research Intern with Tencent.
⋆⋆ Project lead.
⋆⋆⋆ Corresponding author.

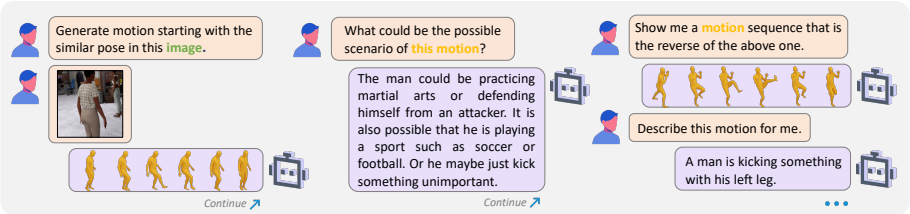


Fig. 1: MotionChain can interpret instructions from multi-turn conversations and generate human motions or textual answers based on text, motion, or image inputs. We provide the conversation results in image-conditioned motion generation (1st column), motion reasoning (second column), motion editing (third column), and motion translation (third column), with each subsequent turn informed by all previous conversations. Left-to-right represents the temporal order.

multi-modal model would have wide-ranging applications in fields like humanoid robotics, virtual assistants, game agents and so on.

Previous research on human motion has explored various tasks, including motion generation [23, 32, 66, 92, 102, 116], motion captioning [20, 24, 32], motion prediction [32, 56, 111, 121], and motion composition [3]. Recent works in text-to-motion [67, 92, 102, 117] have involved pre-trained language models [16, 73] for motion generation. For instance, TEMOS [67] employs BERT [16] text embeddings in an end-to-end transformer architecture, while MDM [92] and MLD [102] both utilize text embeddings from CLIP [73] during the conditional diffusion process. On the other hand, MotionCLIP [91] and TMR [68] focus on modeling the coupled relationship between motion and text description, and MotionGPT [32] introduces a motion-language model that represents human motion and language in one unified vocabulary. However, these above methods treat all tasks as a one-turn conditioned generation, lacking contextual understanding and multi-turn continuous generation abilities. Therefore, we construct a Vision-Motion language model, integrating multi-turn conversations and continuous human motions.

Two crucial challenges need to be addressed in this conversational motion generation. The first challenge is to contextually generate human motion in a continuous manner, resembling the way real humans move. The second challenge is the scarcity of text-motion paired datasets compared to datasets with pairs of image-language [11, 81], image-pose [37, 58, 61, 64] and video-motion [4, 5, 8, 30, 96]. Fortunately, both human motion and language are sequential and can be continuously "written". Building upon this observation, we employ the general vision-language instruction-tuning approach [50, 123] to enable conversational motion generation and question-answering through multi-modal instructions. By integrating image, motion, and language data and encoding them into tokens, the relationship between these three modalities becomes more evident. Therefore, with the advent of large vision-motion and vision-language data, Vision-Motion-language pre-training can enhance the performance of motion-related tasks.

In this study, we introduce MotionChain, a comprehensive framework that integrates vision, motion, and language. MotionChain leverages large-scale vision-language data, vision-motion data, and pre-trained language models' strong language generation abilities to assist in motion-related generation tasks. To enable MotionChain to comprehend and generate human-like motions, we first train a motion-specific vector quantized variational autoencoder (VQ-VAE) model. This model constructs a "motion vocabulary" similar to the English word vocabulary and converts raw motion data into a sequence of motion tokens. To incorporate vision inputs into MotionChain, we then introduce a specialized vision tokenizer that connects a pre-trained vision encoder to the language model. This tokenizer converts image data into visual tokens within the language-motion "words" embedding space. These tokens are then processed by a pre-trained language model [14,74,93,94], which learns the relationship between image, motion and language. To enable conversational generation, we construct a multi-modal motion conversation dataset based on the existing text-motion dataset [23] and vision-motion dataset [5]. We then train the language model using our multi-modal conversation dataset to learn the correlation between the three modalities. Extensive experiments demonstrate that MotionChain achieves state-of-the-art performance in multiple motion-related tasks.

We summarize our contributions as follows: (1) We propose MotionChain, a unified vision-motion-language generative pre-trained model, which performs conversational generation tasks via multi-modal inputs with language models. (2) We introduce a motion composition technique, to generate 3D human motions following the temporal order of instructions. (3) We propose a multi-modal motion conversation benchmark, wherein MotionChain achieves competitive performance across diverse motion tasks.

2 Related Work

Human Motion Modeling. There have been numerous attempts to model the relationship between 3D human motion and multiple modalities including incomplete motion [56, 111, 121], action [3, 25, 40, 66, 99, 102], text [22–24, 32, 55, 67, 82, 92, 117–119], image [19, 21, 33, 114, 115] and video [5, 13, 19, 36, 75]. Text-to-motion is one of the most important motion generation tasks, due to the user-friendly and convenient language input. MDM [92], MotionDiffuse [117] and MLD [102] proposes a diffusion-based generative model [27, 77, 85] to generate motions conditioned on different inputs. TM2T [24] and T2M-GPT [116] investigate a generative framework based on VQ-VAE [76, 95] and generative transformer for motion generation. Motion completion task generates motion conditioning on partial motions, such as classical motion prediction [56,111,121] or motion in-between [92], which generates the intermediate motion while the first and last parts are fixed. TEACH [3] proposes a past-conditioned transformer model that generates motion from a sequence of actions autoregressively. Apart from motion generation, there is also work investigating other modalities of generation from motion. Two statistical models [90] and recurrent networks [70,105]

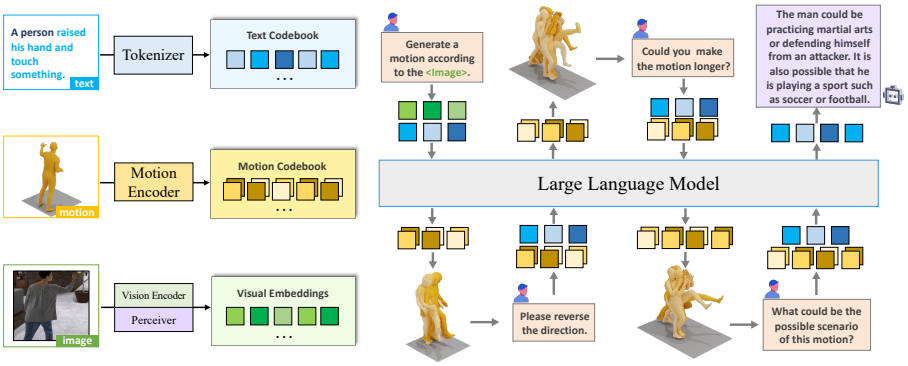


Fig. 2: Method overview: MotionChain consists of a motion tokenizer $\mathcal{V}_{\mathcal{M}}$ (Sec. 3.2), a vision tokenize $\mathcal{V}_{\mathcal{I}}$ (r Sec. 3.2) and a vision-motion-aware language model (Sec. 3.3). By leveraging motion tokens generated by $\mathcal{V}_{\mathcal{M}}$, alongside visual language token embeddings projected by vision tokenizer $\mathcal{V}_{\mathcal{I}}$, and text tokens by text tokenizer, MotionChian achieves a unified learning paradigm for both motion and linguistic data.

are learned in mapping motions to language. TM2T [24] proposed a new motion representation that compresses motions into a short sequence of discrete variables and then uses a neural translation network to build mappings between two modalities. In contrast to the above methods limited to only several tasks, MotionGPT [32] treats human motion as a foreign language and leverages language understanding and zero-shot transfer abilities of pre-trained language models.

Character Control and Animation. Character control involves generating interactive motion sequences based on user instruction signals. One kind of approach [38, 59, 79] is to construct a graph representing transitions between motion clips and plan motion using graph search. Considering the limitations of these graph-based approaches in coarse discreteness, alternative methods like frame blending and concatenation [41], low-dimensional latent space learning [42], motion matching [15] proposed for embedding the task in the feature and [88] do the similar thing through hierarchical setup. Although the control signals for motion control and character animation are different from the instructions in text-to-motion tasks, we still recognize textual commands of conventional human motion generation as a boost for intuitive character control.

Multi-Modal Language Models. In the field of computer vision, there has been a recent surge of interest in multi-modal models that can process text along with other modalities, including images [18, 29, 44, 104], audio [6, 80, 101], and 3D [12,45,54,83,108]. CLIP [73] is an example of such a model, which learns a semantic latent representation that connects images with corresponding language descriptions. While language models have achieved success in various tasks, the development of multi-modal language models capable of handling human motion is still limited. Existing works in computer vision can be broadly categorized into two classes. The first consists of end-to-end trained models explored separately for specific research topics. For example, tasks like vision-language navi-

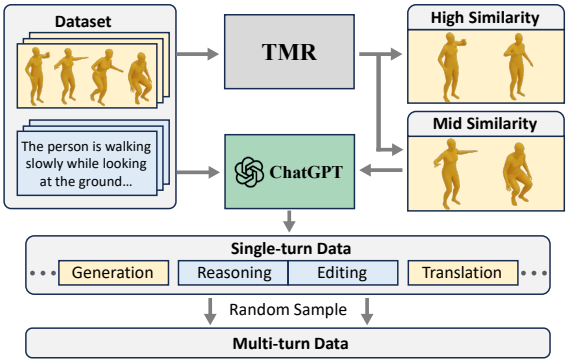


Fig. 3: Data collection overview: Our initial step in collecting the motion reasoning data involves the utilization of human motion captions derived from an existing text-motion dataset. Subsequent to this, the text-motion retrieval model TMR [68] aids in the segmentation of motion pairs into categories based on the similarity between them. With the assistance of ChatGPT, we proceed to craft motion editing task data that correspond to these categorized similarity levels. Incorporating both motion reasoning and editing single-turn tasks, as well as the extensive 14 tasks delineated in [32], we construct a rich multi-modal multi-turn conversation dataset.

gation [2, 7] and Habitat [89] require embodied AI agents to follow natural language instructions and take actions to accomplish goals in visual environments. InstructPix2Pix [7] in image translation enables agents to edit images based on human instructions. The second involves systems that coordinate various models using approaches like LangChain or LLMs [51, 62, 103]. Examples of such systems include Visual ChatGPT [100], X-GPT [125], and MMREACT [110]. While these methods focus on building instruction-following agents, we aim to develop an end-to-end trained multimodal model that can perform conversational motion generation tasks via multi-modal inputs with language models.

3 Methods

To leverage large language data, vision-language data, and vision-motion data for assisting motion-related tasks, we propose a motion-language-vision framework called MotionChain. The framework, as depicted in Fig. 2, consists of **a multi-modal tokenizer** that converts various types of data (text, image, and motion) into discrete tokens (Sec. 3.2), and a **vision-motion-aware Language model** that comprehends information from different modalities and generates corresponding answers based on input instructions (Sec. 3.3). Additionally, to simultaneously understand data from multiple modalities, we employ **a multi-stage training strategy** (Sec. 3.4) for the training of the multi-modal tokenizer and the motion-language-vision framework.

We first introduced the multi-modal tokenizer, which comprises three branches for processing textual, image, and motion inputs. For textual inputs $w^{1:N} = \{w^i\}$ of length N that describes a motion-related question or demand, we employ the SentencePiece model [39] used in previous works [14, 74, 93, 94], which has a vocabulary size of K_t and is trained on a large number of language datasets. The motion branch consists of a motion encoder $\mathcal{E}_{\mathcal{M}}$ that encodes a motion sequence $m^{1:M} = \{x^i\}$ of M frames into L motion tokens $z^{1:L} = \{z^i\}$, where $L = M/l$ and l represents the temporal downsampling rate on motion frames. It also includes a motion decoder $\mathcal{D}_{\mathcal{M}}$ that can decode motion tokens back to human motion $\hat{m}^{1:M}$. The vision branch processes the input image X with a pre-trained CLIP visual encoder and a learnable linear projection that follows it, converted into language token embeddings H_q. Given a textual sentence $w^{1:N}$, a sequence of motion $m^{1:M}$, and an image condition X, all encoded as language tokens, our vision-motion-aware language model is designed to produce an answer comprising L tokens, denoted as $\hat{x}^{1:L} = \{\hat{x}^i\}$. These output tokens can represent either motion sequences $\hat{x}_m^{1:L}$ or textual descriptions $\hat{x}_t^{1:L}$, which integrate both human motion $\hat{m}^{1:M}$, and text $\hat{w}^{1:L}$ within the given context.

3.1 Data Collection

With the emergence of text-conditioned motion generation tasks, datasets like KIT [69], BABEL [71], HumanML3D [23] and the more recent Motion-X [49] have been developed. However, these datasets predominantly offer text labels as simple action phrases or captions. Building upon these foundations, MotionGPT [32] introduces an instruction-based motion-language dataset that encapsulates 14 core tasks, including motion prediction, translation, and editing, through thousands of instruction templates in a unified format. Despite this advancement, MotionGPT's data lack a deep engagement with the nuances of human motion analysis and are limited to single-turn generation tasks without incorporating contextual memory. Inspired by the recent success of GPT models across text-annotation tasks [17], image-annotation tasks [50], 3D-annotation tasks [28], we propose a data collection methodology integrates the capabilities of existing LLMs like ChatGPT [62], with the text-motion retrieval model TMR [68] to facilitate motion conversation data collection. In addition to the 14 motion-related tasks in MotionGPT [32], we introduce tasks centered around motion reasoning and motion editing, leveraging contextual insights for a deeper motion analysis.

Utilizing ChatGPT [62], we initiate the collection of motion reasoning data using human motion captions from the text-motion dataset [23], starting with manually designed example queries that explore the contextual scenarios surrounding motions, possible preceding or succeeding actions, the subjects' roles, and the tools or equipment involved, etc. Following this, we employ TMR [68] for categorizing motions from the dataset into varying similarity levels. For medium-similarity motion pairs, we utilize ChatGPT [62] to generate motion editing directives that enable the transformation of one motion to another. For motions of

high similarity, we manually devise tasks aimed at editing their lengths, further enriching the dataset's versatility and analytical scope.

After the collection of single-turn generation tasks, we progress to develop multi-turn conversation data. This involves the deliberate association of initial motion generation tasks with a variety of follow-up tasks randomly chosen among motion translation, reasoning, editing, etc. Following [122] we construct our conversation data in a structured format, as depicted below:

$X_{\text{system-message}}$
USER: X_v X_s^1 ASSISTANT: X_a^1 </s>
USER: X_v X_s^2 ASSISTANT: X_a^2 </s>
USER: X_v X_s^3 ASSISTANT: X_a^3 </s> ...

Where X_v is defined as the vision language token embeddings, processed via the visual tokenizer. X_s^i and X_a^i are used to denote the source inputs and target answers for each round i, respectively. Both sets of tokens originate from the integrated motion-language vocabulary V, which includes motions, texts, or a blend thereof. The dataset exhibits variability in the number of generation turns up to 10; for the sake of clarity, we present only three examples herein. MotionChain is trained to predict answers, incorporating a learning mechanism that determines whether to stop generation by outputting end of sentence flag </s> based on the current instruction and all preceding questions and answers. In the computation of the loss, as defined in Eq. (5), only the green tokens are utilized.

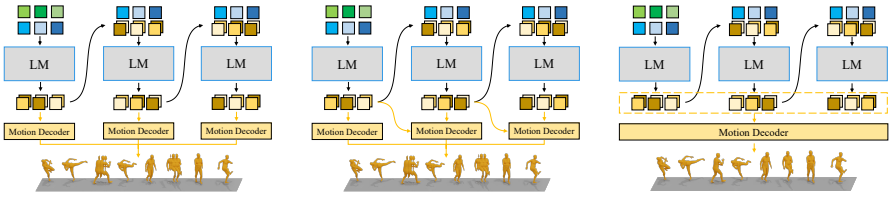
3.2 Multi-modal Tokenizer

Motion tokenizer, denoted as $\mathcal{V}_{\mathcal{M}}$, is based on the architecture of Vector Quantized Variational Autoencoders (VQ-VAE) utilized in previous studies [22,24,32, 84,95,98,106,109,116]. Once pre-trained, it can represent motion using discrete tokens, facilitating the integration of motion and language. The Motion tokenizer consists of a motion encoder $\mathcal{E}_{\mathcal{M}}$ and a motion decoder $\mathcal{D}_{\mathcal{M}}$. Initially, the motion encoder $\mathcal{E}$ applies 1D convolutions to the motion features $m^{1:M}$ along the temporal dimension to obtain latent vectors $\hat{z}^{1:L} = \mathcal{E}_{\mathcal{M}}(m^{1:M})$. Subsequently, the latent vectors $\hat{z}$ are quantized and transformed into a collection of codebook entries z. The learnable codebook $Z = \{z_i\}_{i=1}^K \subset \mathbb{R}^d$ comprises K latent embedding vectors, each with a dimension of d. The quantization process $Q(\cdot)$ replaces each row vector $\hat{z}$ with its nearest codebook entry z_k in Z, which can be expressed as:

$$z_i = Q(\hat{z}^i) := \arg\min_{z_k \in Z} \|\hat{z}_i - z_k\|_2 . \tag{1}$$

We assign s^i as the index number of motion tokens $z^{1:L}$, so motion tokens $z^{1:L}$ can be represented as a sequence of indices $s^{1:L} = \{s^i\}_{i=1}^L$. The motion decoder $\mathcal{D}_{\mathcal{M}}$ can project $z^{1:L} = \{z^i\}_{i=1}^L$ back to the raw motion space, resulting in the motion $\hat{m}^{1:M}$ with M frames. Following [22,24,32,98,116], we adopt three distinct loss functions when training the motion tokenizer:

$$\mathcal{L}_{\mathcal{V}} = \mathcal{L}_r + \mathcal{L}_e + \mathcal{L}_c \tag{2}$$



(a) Independent **(b)** Past-condition **(c)** Tokens-joint

Fig. 4: Motion Composition Variants: We illustrate the baselines for motion composition during multi-turn motion generation (a). independent decoding each turn (b). separate decoding conditioned on the last few tokens from the prior turn (c). decoding with joint motion tokens. Green tokens stand for image condition, blue tokens stand for textual instruction, and orange tokens stand for human motions.

where $\mathcal{L}_r$ denotes reconstruction loss, $\mathcal{L}_e$ denotes the embedding loss, and $\mathcal{L}_c$ denotes commitment loss.

During multi-turn motion generation, the motion continuity between turns is achieved through our motion decoder, which links the motion of the current turn with that of the preceding ones. Taking the composition of two motions as an example: we concatenate the past motion tokens, denoted as $z_p^{1:L_p}$, with the tokens representing the current motion, $z_c^{1:L_c}$. This concatenated sequence of tokens is subsequently decoded into a comprehensive set of continuous motion features, represented as $m_{\text{whole}}^{1:M_{\text{whole}}}$, as depicted below:

$$z_{\text{whole}}^{1:(L_p+L_c)} = [z_p^{1:L_p}, z_c^{1:L_c}]. \tag{3}$$

Similarly, this framework is adept at executing composition tasks involving an array of motions. The comparison results in Tab. 3 demonstrate that our motion tokenizer could effectively perform motion composition tasks.

Visual Tokenizer accepts images X_I as inputs. We employ the CLIP visual encoder that is pre-trained on image-text pairs to derive visual feature Z_I. These features are then projected into language token embeddings X_v via a linear layer like previous work [50], maintaining consistency in the dimensionality with the language model's word embedding space.

3.3 Motion-aware Language Model

Language models such as Llama [93, 94] and T5 [14, 74] employ the SentencePiece [39] model to encode textual inputs into WordPiece tokens, utilizing a K_t word piece vocabulary. Unlike prior text-to-motion [24,102,116,118] and motion-to-text [24] methods that process text and motion separately, we merge the text vocabulary $V_t = \{v_t^i\}_{i=1}^{K_t}$ with the motion vocabulary $V_m = \{v_m^i\}_{i=1}^{K_m}$, maintaining the motion tokenizer's codebook Z order and including special tokens for boundary demarcation. This creates a unified vocabulary $V = \{V_t, V_m\}$, enabling the formulation of motion-centric tasks in a universal template, where inputs and

outputs share the same vocabulary. For visual input, our visual tokenizer converts images into visual token embeddings X_v, aligning with the language model [39, 62, 74, 93] token space for integrated representation.

For single conditioned generation tasks, our input comprises a sequence of N length tokens $X_s = \{{x_s}^i\}_{i=1}^N$, where $x_s \in \{V_t, V_m\}$ representing either text, motion, or a combination thereof, drawn from the unified vocabularies. In cases involving image inputs, visual tokens X_v are interspersed at the beginning of the source tokens sequence, forming $[X_v, X_s]$. Subsequent interaction rounds generate target answer tokens X_a. To facilitate iterative result generation and content retention, our framework generates multi-turn conversation data $(X_v, X_s^1, X_a^1, X_s^2, X_a^2, \cdots, X_s^T, X_a^T)$, with T indicating the total turn count. Notably, visual tokens are consistently placed at the forefront of the initial turn's source tokens. The processing sequence is organized such that to predict target answer tokens autoregressively, as shown in Fig. 2. Source tokens are processed by the transformer to predict the next token's probability distribution, formulated as:

$$p_\theta(X_a \mid X_v, X_s) = \prod_i p_\theta\left(x_a^i \mid X_v, X_{s,<i}, X_{a,<i}\right) \tag{4}$$

with θ indicating trainable parameters, and $X_{s,<i}, X_{a,<i}$ representing the sequences of source and preceding target tokens. The training objective is maximizing the log-likelihood of distribution:

$$\mathcal{L}_{LM} = -\sum_{i=0}^{L_t-1} \log p_\theta\left(x_a^i \mid X_v, X_{s,<i}, X_{a,<i}\right). \tag{5}$$

By optimizing this objective, MotionChain captures the complex interrelations among images, motion, and text, facilitating accurate target "word" generation.

During the inference phase, target tokens are recursively sampled from the model's predicted distribution $p_\theta\left(\hat{x_a}^i \mid X_v, X_s, \hat{X}_{a,<i}\right)$, ceasing with the appearance of a special end token. This strategy facilitates a step-by-step target sequence generation, where each token's probability is conditioned on all previous turns' sources and targets and current source input.

3.4 Training Strategy

To facilitate the integration of image and motion comprehension within the language modeling context, we adopt a 3-stage training strategy. (1) The initial stage involves pre-training the motion tokenizer on a corpus of human motion data, in line with [32]. This process establishes the motion vocabulary V_m, which serves as a foundation for encoding human motions as a series of discrete tokens. (2) Subsequently, the motion tokenizer remains frozen while we connect the visual tokenizer to the language model framework. This integration is supported by a suite of supervised objectives, including text-to-motion, motion-to-text, and image-based motion generation, aiming to learn the intricate relationships between images, motion, and language. (3) The final stage involves instruction

tuning, and refines the model's capabilities through the application of prompt-based instructions. These instructions are framed within multi-turn conversation sequences, as detailed in Sec. 3.3, to expanded range of motion-related tasks.

Training of Motion Tokenizer. The initial step involves training the motion tokenizer, guided by the loss objective in Equation 2. This stage enables the tokenizer to represent human motion sequences $\hat{x}^{1:L}$ as discrete motion tokens, a key step for merging motion data with textual information seamlessly. Once optimized, the motion tokenizer remains frozen.

Motion-language Pre-training Stage. Leveraging recent developments in language modeling [14, 74, 93, 94, 122] pre-trained on natural language datasets and then fine-tuned with instruction-based phrasing [14, 62]. To augment the model's ability to discern relationships between images and human motions, we first pre-train our MotionChain using a mix of language, image, and motion datasets. Following the stage 1 training of the motion tokenizer, we have established a unified motion-language vocabulary $V = V_t, V_m$, capable of representing motions in discrete token form. Moreover, we maintain the visual encoder's weights in the visual tokenizer as fixed, while the linear projection weight W is jointly optimized with the language model. During this stage, the model undertakes three fundamental single-turn modality translation tasks: text-to-motion, motion-to-text, and image-conditioned motion generation, as outlined in Sec. 3.1. The primary objective is to maximize the likelihood of the model according to the loss function specified in Eq. (5), thereby letting the model understand the relationship between language, vision conditions, and motions.

Instruction Tuning Stage. As described in Sec. 3.1, we construct a multi-modal, multi-task, and multi-turn motion conversation dataset by augmenting existing text-to-motion [23] and human mesh reconstruction datasets [5] with targeted instructional prompts and leverage the capabilities of LLMs [62] and the text-motion retrieval model [68] for motion reasoning and editing tasks. The efficacy of instruction tuning, as evidenced across language models [14, 50, 62, 122], is well-established, yielding enhancements in model performance across a wide range of tasks. After instruction tuning, MotionChain can handle more motion-related tasks including the proficient handling of previously unseen tasks

4 Experiments

We evaluate the proposed MotionChain encompasses comprehensive comparisons across both one-turn motion-related tasks and multi-turn motion generation tasks. Firstly, we provide details of the dataset settings, evaluation criteria, and implementation details as specified in Sec. 4.1. Subsequently, comparative analyses are presented, focusing on the motion reasoning task (Sec. 4.2) and the temporal motion composition task (Sec. 4.3). In Sec. 4.4, we evaluate the choice of motion composition technique and different architectures of vision tokenizer.

4.1 Experimental Setup

Datasets. For one-turn motion reasoning tasks, the study employs our proposed multi-modal multi-turn conversation dataset upon HumanML3D [23] with 44,970 sequence-level textual descriptions for 14,616 motion sequences obtained from AMASS [57] and HumanAct12 [25]. The datasets are divided into training, testing, and validation sets with a ratio of 0.8 : 0.15 : 0.05. To evaluate the multi-turn motion generation task, we focus on BABEL [71] that provides textual descriptions for the motions in the AMASS [57] with annotated segments that overlap in each sequence, which allows evaluating generation of a sequence of motion or actions. We adopt the processed text labels by [3] and motion representation of HumanML3D [23] which combines joint velocities, positions, and rotations. Following [3] we consider pairs of actions for simplicity but MotionChain applies to a sequence of actions or motion of arbitrary length. For the image-conditioned motion generation task, we mainly focus on BEDLAM [5], a large synthetic dataset of realistic moving 3D humans containing more than 200 subjects and 380K frames video and motion pair.

Evaluation Metrics are summarized as four parts. (1) Motion quality: We adopt Frechet Inception Distance (FID) as the primary metric, FID quantifies the divergence in feature distributions between generated and actual motion sequences. Utilizing feature extractors from prior studies [23,49,68], FID measures the distance of feature distributions between the generated and real motions. Following [5,32,33,65,102], we also adopt MPJPE, PA-MPJPE to measure global and local errors in millimeters and ACCL for acceleration errors, to evaluate the quality of the reconstructed motions. (2) Motion Diversity: Utilizing the Diversity (DIV) metric, we calculate variance across motion features to evaluate generation diversity. (3) Text matching: The precision of text-to-motion matches is quantified by the R Precision metric, based on the feature evaluator [23,49,68], and includes an analysis of Top 1/2/3 retrieval accuracy. The Multi-modal Distance (MM Dist) quantifies the semantic gap between motions and texts. (4) Linguistic quality: We follow [24] utilizing linguistic metrics from natural language studies, including BLUE [63], Rouge [48], Cider [97], and BertScore [120] to evaluate the quality of generated motion captions. More detailed benchmark information is provided in the supplementary materials.

Implementation Details. We set the codebook of the motion tokenizer as $K \in \mathbb{R}^{512\times1024}$ for most experiments. The motion encoder, denoted as $\mathcal{E}_{\mathcal{M}}$, integrates a temporal downsampling rate, $l = 4$. Our vision tokenizer incorporates a frozen Vision Transformer (ViT-L/14) [73] as visual encoder for most experiments. Additionally, for comprehensive ablation studies, we explored the use of both a frozen vision encoder and a Q-former from BLIP-2 [43] as a vision tokenizer. We mainly utilize Flan-T5-base [14] as the underlying architecture for our language model. Moreover, all our models employ the AdamW [53] optimizer with $[\beta_1, \beta_2] = [0.9, 0.99]$ for training. The motion tokenizers are trained to utilize a 10^{-4} learning rate employing cosine annealing scheduler and a 256 mini-batch size. Our language models based on Flan-T5-base [14] have a 10^{-4} learning rate with cosine annealing scheduler and 16 mini-batch sizes in both the

Table 1: Comparison of motion reasoning on the test set of our conversation dataset. Our proposed MotionChain is fine-tuned on motion reasoning tasks while other methods' results are generated by their pre-trained weight. $Length_{avg}$ represents the average words in generated answers to all questions. We adopt metrics commonly used in natural language processing tasks for evaluation.

Methods	Params	$\text{Length}_{\text{avg}}$	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	BertScore↑
Flan-t5-base [14]	250M	8.34	4.64	1.78	15.32	15.93	3.45
Flan-t5-large [14]	780M	11.95	12.18	4.83	22.81	15.02	14.19
Flan-t5-xl [14]	3B	9.09	8.54	4.01	24.89	15.03	18.34
Llama-2-7b [94]	7B	130.84	11.12	3.67	19.14	1.04	6.81
Vicuna-1.5-7b [122]	7B	71.49	19.27	7.39	25.75	5.44	19.05
Vicuna-1.5-13b [122]	13B	84.74	17.20	6.53	24.18	7.77	18.00
MotionChain (Ours)	573M	22.17	**37.92**	**19.19**	**38.05**	**24.53**	**32.24**

pre-train stage and the instruction tuning stage. The motion tokenizer undergoes 10000 epochs of training, while the language model undergoes 500 epochs during the pre-train stage and another 50 epochs during the instruction tuning stage. Most models are trained on 8 Tesla V100 GPUs.

4.2 Comparisons on Motion Reasoning.

In Sec. 3.1, we introduce a multi-modal motion conversation dataset, enriched with motion reasoning data facilitated by ChatGPT [62]. This task evaluates the model's reasoning capabilities with motion reasoning tasks, where a motion sequence or its corresponding textual descriptions serve as inputs. Our evaluation compares our MotionChain, which integrates motion perception, against contemporary Large Language Models (LLMs) that possess solely textual processing capabilities. The compared LLMs are assessed using their original pre-trained weight. Results in Tab. 1, illustrate that MotionChain exhibits superior motion reasoning proficiency, benefiting from its integrated motion perception.

4.3 Comparisons on Temporal Composition.

The temporal motion composition task involves generating a continuous motion sequence from two actions in a time series. We conducted our experiments following the settings in TEACH [3] and used the Amass [57] subset BABEL [71] validation set. Additionally, we processed the motion in AMASS [57] into the format proposed by HumanML3D [23] and trained our MotionChain on the action-to-motion task. To compare with TEACH, we initially used an officially provided pre-trained model to sample motion on the validation set 20 times. Subsequently, we post-processed their motion into the HumanML3D format, represented in SMPL [52]. The performance of our MotionChain is summarized in Table 2. As evaluating generative models quantitatively is challenging, we also provide qualitative comparisons in the supplementary materials.

Table 2: Comparison of temporal motion composition on Babel [71]. We evaluate the state-of-the-art motion temporal composition method Teach [3] under the 95 % confidence interval from 20 times running. (*cf*. Sec. 4.1 for notations.)

Methods	Diversity	MPJPE↓	PA-MPJPE↓	ACCL↓
Real	$15.74^{\pm.149}$	-	-	-
Teach [3]	$27.11^{\pm.159}$	$979.21^{\pm.215}$	$933.32^{\pm.254}$	$23.02^{\pm.018}$
MotionChain (Ours)	$43.25^{\pm.159}$	$\mathbf{276.05}^{\pm6.72}$	$\mathbf{53.72}^{\pm.580}$	$\mathbf{7.11}^{\pm0.100}$

Table 3: Evaluation of motion composition methods on HumanML3D [23]. Here *Independent*, *Past-condition*, and *Tokens-joint* stand for different motion composition varients during multi-turn motion conversation, as illustrated in Fig. 4.

Method	MPJPE↓	PA-MPJPE↓	ACCL↓	Diversity
Independent	350.79	102.97	11.40	6.47
Past-condition	232.46	46.15	6.18	6.01
Tokens-joint	**108.77**	**18.85**	**2.26**	5.56

4.4 Ablation Studies

MotionChain enables multi-modal motion conversation using two main techniques. The first technique involves generating a smooth sequence of motions by concatenating motion tokens which are then decoded back to motion by motion decoder $\mathcal{D}_{\mathcal{M}}$. The second technique involves processing multi-modal visual input through a vision tokenizer, which consists of a frozen vision encoder and a trainable linear projection. To evaluate the effectiveness of these two designs, we compare them with other variants. For a more comprehensive analysis, detailed ablation studies can be found in the supplementary materials.

Motion Composition Mechanism Apart from the jointly token concatenating mechanism, we also evaluate the performance of temporal motion composition through the other motion temporal composition variants Motion-cat: concatenating the motion in final motion level rather than token level. Experimental results in Tab. 3 show that jointly concatenating motion tokens achieved remarkable performance compared to the other variants. For further information regarding the implementation of the aforementioned vision tokenizer, please refer to the supplementary materials.

Image Tokenizeer Architecture. MotionChain connects the frozen vision encoder to the language model through a linear layer. However, previous vision-language [1, 43] works also demonstrate the effectiveness of other kinds of visual-aligning modules. Here we consider the other two vision tokenizer variants: (a) inspired by [1, 10, 31], we introduce a perceiver module that incorporates a transformer receiving a predefined number of latent input queries. These queries cross-attend to the visual features, enabling effective information exchange. (b) We directly adopt the pre-trained Q-former from BLIP-2 [43] to align visual

Table 4: Evaluation of vision tokenizer architecture on Bedlam [49]. We implement three different architectures, including Q-former, Perceiver, and Linear. We evaluate these results with the metrics in motion reconstruction. Additional information regarding the implementation is in the supplementary materials. (*cf.* Tab. 2 for notations.)

Architecture	First-frame		Last-frame	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Q-former	195.49	86.56	134.73	57.17
Perceiver	185.61	99.21	134.89	57.58
Linear	**144.37**	**76.48**	**133.73**	**56.73**

inputs with the language model. We evaluate the different architectures under the single human image as the first frame condition and the last frame condition separately. Experimental results in Tab. 4 show that a lightweight linear projection is sufficient for comprehending the human pose from visual input. Additional details about the implementation of the above vision tokenizer can be found in the supplements.

5 Conclusion and Limitation

Limitation. As the trial to explore conversational human motion generation with visual language models, the proposed MotionChain still has limitations as follows. MotionChain utilizes indeterministic generative models, similar to other language models, but other traditional or neural motion controllers [86, 87] are mostly deterministic and sensitive to control signals. Besides, our method can only generate motion on articulated human bodies, excluding many other human parts such as faces [9, 34, 72] and hands [46, 46, 47, 47, 78]. Although we utilize vision, language, and motion as multimodal conditional inputs akin to human perception, MotionChain is still restricted to the collision signals for human-object and human-scene interactions [35, 82, 112].

Conclusion. We summarize the proposed MotionChain as a conversational human motion controller to generate continuous and long-term human motion through multimodal prompts. Compared to these one-turn motion generation methods [32, 92, 117], our MotionChain produces more contextually rich generation and can achieve the step-by-step process of human task execution for humanoid robotics and game agents. By leveraging large-scale language, vision-language, and vision-motion data to assist motion-related generation tasks, MotionChain thus comprehends each instruction in multi-turn conversation and generates human motions followed by these prompts. Extensive experiments validate the efficacy of MotionChain, demonstrating state-of-the-art performance in conversational motion generation, as well as more intuitive manners of controlling and interacting with virtual humans.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62101137), National Key Research and Development Program of China (No. 2022ZD0160101), Shanghai Natural Science Foundation (No. 23ZR1402900), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103). The computations in this research were performed using the CFFF platform of Fudan University.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3674–3683 (2018)
3. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action compositions for 3d humans. In: International Conference on 3D Vision (3DV) (September 2022)
4. Bazavan, E.G., Zanfir, A., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Hspace: Synthetic parametric humans animated in complex environments. arXiv preprint arXiv:2112.12867 (2021)
5. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8726–8737 (2023)
6. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., et al.: Audiolm: a language modeling approach to audio generation. IEEE/ACM transactions on audio, speech, and language processing **31**, 2523–2533 (2023)
7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
8. Cai, Z., Zhang, M., Ren, J., Wei, C., Ren, D., Lin, Z., Zhao, H., Yang, L., Loy, C.C., Liu, Z.: Playing for 3d human recovery. arXiv preprint arXiv:2110.07588 (2021)
9. Cao, X., Chen, Z., Chen, A., Chen, X., Li, S., Yu, J.: Sparse photometric 3d face reconstruction guided by morphable models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4635–4644 (2018)
10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
11. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)

12. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26428–26438 (2024)
13. Choudhury, R., Kitani, K.M., Jeni, L.A.: Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14750–14760 (2023)
14. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
15. Clavet, S.: Motion matching and the road to next-gen animation. In: Proc. of GDC. vol. 2, p. 9 (2016)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
17. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056 (2023)
18. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. arXiv preprint arXiv:2305.05665 (2023)
19. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. arXiv preprint arXiv:2305.20091 (2023)
20. Goutsu, Y., Inamura, T.: Linguistic descriptions of human motion with generative adversarial seq2seq learning. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 4281–4287. IEEE (2021)
21. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10884–10894 (2019)
22. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. arXiv preprint arXiv:2312.00063 (2023)
23. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
24. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
25. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
26. Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., et al.: Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. arXiv preprint arXiv:2309.00615 (2023)
27. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
28. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. arXiv preprint arXiv:2307.12981 (2023)

29. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al.: Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045 (2023)
30. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
31. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)
32. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. arXiv preprint arXiv:2306.14795 (2023)
33. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018)
34. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
35. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
36. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
37. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: Spec: Seeing people in the wild with an estimated camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11035–11045 (2021)
38. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 723–732 (2023)
39. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)
40. Lee, T., Moon, G., Lee, K.M.: Multiact: Long-term 3d human motion generation from multiple action labels. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1231–1239 (2023)
41. Lee, Y., Wampler, K., Bernstein, G., Popović, J., Popović, Z.: Motion fields for interactive character locomotion. In: ACM SIGGRAPH Asia 2010 papers, pp. 1–8 (2010)
42. Levine, S., Wang, J.M., Haraux, A., Popović, Z., Koltun, V.: Continuous character control with low-dimensional embeddings. ACM Transactions on Graphics (TOG) **31**(4), 1–10 (2012)
43. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
44. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
45. Li, M., Chen, X., Zhang, C., Chen, S., Zhu, H., Yin, F., Yu, G., Chen, T.: M3dbench: Let's instruct large models with multi-modal 3d prompts. arXiv preprint arXiv:2312.10763 (2023)

46. Li, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Piano: A parametric hand bone model from magnetic resonance imaging. arXiv preprint arXiv:2106.10893 (2021)
47. Li, Y., Zhang, L., Qiu, Z., Jiang, Y., Li, N., Ma, Y., Zhang, Y., Xu, L., Yu, J.: Nimble: a non-rigid hand model with bones and muscles. ACM Transactions on Graphics (TOG) **41**(4), 1–16 (2022)
48. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
49. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. arXiv preprint arXiv:2307.00818 (2023)
50. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
51. Liu, X., Yan, H., Zhang, S., An, C., Qiu, X., Lin, D.: Scaling laws of rope-based extrapolation. arXiv preprint arXiv:2310.05209 (2023)
52. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (Oct 2015)
53. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
54. Lu, C., Yin, F., Chen, X., Liu, W., Chen, T., Yu, G., Fan, J.: A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7557–7567 (2023)
55. Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Humantomato: Text-aligned whole-body motion generation. arXiv preprint arXiv:2310.12978 (2023)
56. Ma, H., Li, J., Hosseini, R., Tomizuka, M., Choi, C.: Multi-objective diverse human motion prediction with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8161–8171 (2022)
57. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
58. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: International Conference on 3D Vision (3DV) (2017). https://doi.org/10.1109/3dv.2017.00064, http://gvv.mpi-inf.mpg.de/3dhp_dataset
59. Min, J., Chai, J.: Motion graphs++ a compact generative model for semantic motion analysis and synthesis. ACM Transactions on Graphics (TOG) **31**(6), 1–12 (2012)
60. OpenAI: Gpt-4 technical report (2023)
61. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems **24** (2011)
62. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
63. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)

64. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: Agora: Avatars in geography optimized for regression analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13468–13478 (2021)
65. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
66. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: International Conference on Computer Vision (ICCV) (2021)
67. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV) (2022)
68. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. arXiv preprint arXiv:2305.00976 (2023)
69. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big Data **4**(4), 236–252 (dec 2016). https://doi.org/10.1089/big.2016.0028, http://dx.doi.org/10.1089/big.2016.0028
70. Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Robotics and Autonomous Systems **109**, 13–26 (2018)
71. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 722–731 (Jun 2021)
72. Qiu, Z., Li, Y., He, D., Zhang, Q., Zhang, L., Zhang, Y., Wang, J., Xu, L., Wang, X., Zhang, Y., et al.: Sculptor: Skeleton-consistent face creation using a learned parametric generator. ACM Transactions on Graphics (TOG) **41**(6), 1–17 (2022)
73. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
74. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
75. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3d appearance, location and pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2740–2749 (2022)
76. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
77. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022), https://github.com/CompVis/latent-diffusionhttps://arxiv.org/abs/2112.10752
78. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
79. Rose, C., Cohen, M.F., Bodenheimer, B.: Verbs and adverbs: Multidimensional motion interpolation. IEEE Computer Graphics and Applications **18**(5), 32–40 (1998)

80. Rubenstein, P.K., Asawaroengchai, C., Nguyen, D.D., Bapna, A., Borsos, Z., Quitry, F.d.C., Chen, P., Badawy, D.E., Han, W., Kharitonov, E., et al.: Audiopalm: A large language model that can speak and listen. arXiv preprint arXiv:2306.12925 (2023)
81. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
82. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
83. Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., Nießner, M.: Meshgpt: Generating triangle meshes with decoder-only transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19615–19625 (2024)
84. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11050–11059 (2022)
85. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
86. Starke, S., Mason, I., Komura, T.: Deepphase: periodic autoencoders for learning motion phase manifolds. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022)
87. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209–1 (2019)
88. Starke, S., Zhao, Y., Zinno, F., Komura, T.: Neural animation layering for synthesizing martial arts movements. ACM Transactions on Graphics (TOG) **40**(4), 1–16 (2021)
89. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D.S., Maksymets, O., et al.: Habitat 2.0: Training home assistants to rearrange their habitat. Advances in Neural Information Processing Systems **34**, 251–266 (2021)
90. Takano, W., Nakamura, Y.: Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. The International Journal of Robotics Research **34**(10), 1314–1328 (2015)
91. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 358–374. Springer (2022)
92. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A.H., Cohen-Or, D.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
93. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
94. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
95. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

96. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Computer Vision and Pattern Recognition (CVPR) (2017)
97. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
98. Wang, C.: T2m-hifigpt: Generating high quality human motion from textual descriptions with residual discrete representations. arXiv preprint arXiv:2312.10628 (2023)
99. Wang, W., Zhe, X., Chen, H., Kang, D., Li, T., Chen, R., Bao, L.: Neural marionette: A transformer-based multi-action human motion synthesis system. arXiv preprint arXiv:2209.13204 (2022)
100. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
101. Wu, Y.C., Gebru, I.D., Marković, D., Richard, A.: Audiodec: An open-source streaming high-fidelity neural audio codec. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
102. Xin, C., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, J., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
103. Xiong, W., Liu, J., Molybog, I., Zhang, H., Bhargava, P., Hou, R., Martin, L., Rungta, R., Sankararaman, K.A., Oguz, B., et al.: Effective long-context scaling of foundation models. arXiv preprint arXiv:2309.16039 (2023)
104. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021)
105. Yamada, T., Matsunaga, H., Ogata, T.: Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Robotics and Automation Letters **3**(4), 3441–3448 (2018)
106. Yao, H., Song, Z., Zhou, Y., Ao, T., Chen, B., Liu, L.: Moconvq: Unified physics-based motion control via scalable discrete representations (2023)
107. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
108. Yin, F., Chen, X., Zhang, C., Jiang, B., Zhao, Z., Fan, J., Yu, G., Li, T., Chen, T.: Shapegpt: 3d shape generation with a unified multi-modal language model. arXiv preprint arXiv:2311.17618 (2023)
109. Yin, F., Liu, W., Huang, Z., Cheng, P., Chen, T., Yu, G.: Coordinates are not lonely-codebook prior helps implicit neural 3d representations. Advances in Neural Information Processing Systems **35**, 12705–12717 (2022)
110. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 (2022)
111. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 346–364. Springer (2020)

112. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: PhysDiff: Physics-guided human motion diffusion model. In: IEEE International Conference on Computer Vision (ICCV) (October 2023)
113. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
114. Zhang, H., Cao, J., Lu, G., Ouyang, W., Sun, Z.: Danet: Decompose-and-aggregate network for 3d human shape and pose estimation. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 935–944 (2019)
115. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
116. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: 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) (2023)
117. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
118. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023)
119. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10188–10198. IEEE (2023)
120. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
121. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3372–3382 (2021)
122. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023)
123. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al.: Lima: Less is more for alignment. Advances in Neural Information Processing Systems **36** (2024)
124. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
125. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)