Dense Motion Captioning

Shiyao Xu[1] Benedetta Liberatori[1] Gül Varol[2] Paolo Rota[1]
[1]University of Trento [2]LIGM, Ecole des Ponts, IP Paris, Univ Gustave Eiffel, CNRS
shiyao.xu@unitn.it
xusy2333.com/demo

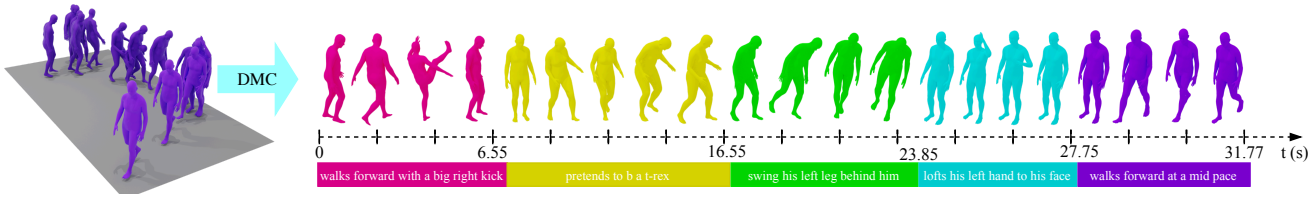

Figure 1. **Dense Motion Captioning (DMC).** We present DMC, a task that localizes and generates detailed segment-level captions with accurate temporal boundaries in 3D human motion sequences. To support this task, we construct CompMo, the first large-scale 3D motion-language dataset providing dense captions for multiple temporal segments within each motion sequence. Each sequence contains between 2 and 10 atomic actions, and every action is annotated with precise timestamps and a descriptive caption.

Abstract

Recent advances in 3D human motion and language integration have primarily focused on text-to-motion generation, leaving the task of motion understanding relatively unexplored. We introduce Dense Motion Captioning, a novel task that aims to temporally localize and caption actions within 3D human motion sequences. Current datasets fall short in providing detailed temporal annotations and predominantly consist of short sequences featuring few actions. To overcome these limitations, we present the Complex Motion Dataset (CompMo), the first large-scale dataset featuring richly annotated, complex motion sequences with precise temporal boundaries. Built through a carefully designed data generation pipeline, CompMo includes 60,000 motion sequences, each composed of multiple actions ranging from at least two to ten, accurately annotated with their temporal extents. We further present DEMO, a model that integrates a large language model with a simple motion adapter, trained to generate dense, temporally grounded captions. Our experiments show that DEMO substantially outperforms existing methods on CompMo as well as on adapted benchmarks, establishing a robust baseline for future research in 3D motion understanding and captioning.

1. Introduction

Recently, there has been a growing interest in integrating 3D human motion and language modalities. Most progress in this area has focused on text-to-motion generation [12, 16, 27, 41, 43, 46], which involves synthesizing 3D human movements from natural language descriptions, and motion editing [1, 12, 13], where existing motion sequences are modified according to textual instructions. These tasks have advanced rapidly, driven by the development of datasets that pair 3D human motions with language descriptions [10, 21, 30, 31].

In contrast, 3D human motion understanding remains in its infancy. While some recent works have begun to explore this direction, most efforts focus on relatively simple tasks such as motion-to-text retrieval [3, 7, 28] or captioning of short, isolated motion sequences [11, 16, 43, 52]. Understanding longer and more complex motion sequences with temporal precision is crucial for applications that require a detailed understanding of human activities. For example, by lifting 2D videos into 3D motion representations and generating temporally grounded descriptions from this data, we can develop systems that go beyond traditional video analysis. This approach allows for a more accurate, body-centric understanding, especially in situations where subtle nuances of motion are crucial.

Motivated by this, we introduce Dense Motion Captioning (DMC) as a new task and experimental setting, which involves detecting all semantically meaningful actions in a motion sequence, captioning them, and determining their precise start and end times. Unlike traditional single-motion captioning, this task involves parsing a continuous stream of motion and segmenting it into temporally localized action units.

A major limitation of existing benchmarks is their lack of complex motion sequences as well as precise annotations. Most available datasets contain only isolated actions or a few simple actions concatenated together, or suffer from noisy

annotations, where the descriptions or labels are fragmented and lack consistency. In our preliminary experiment on the HumanML3D dataset (see Sec. 3.1), we aim to assess whether current motion captioning models can maintain their performance when handling longer motion sequences containing more than a single action. Our findings indicate a notable performance drop under these conditions. To address this limitation, we introduce the Complex Motion Dataset (CompMo), a large-scale dataset specifically designed for dense motion captioning. As illustrated in Fig. 1, it features extended motion clips with multiple actions. Each action is annotated with a detailed caption and temporal boundaries. Alongside the dataset, we design Dense Motion Captioning Model (DEMO), a strong baseline that generates detailed, temporally aligned captions from long and complex 3D motion sequences. DEMO is composed of a Large Language Model (LLM) and a simple motion adapter. It is trained in two stages: first, to align motion and language modalities, and second, to finetune the model for dense caption generation. We evaluate it on CompMo and existing motion-language datasets repurposed for the DMC setting, establishing the first comprehensive benchmark for this task.

In summary, this work makes three main contributions. First, we introduce DMC, a novel task which aims to generate sequences of textual descriptions for complex motions, with temporal boundaries. Second, we present CompMo, a large-scale dataset specifically curated for this task, featuring rich annotations that capture diverse and intricate human motions across multiple scenarios. Finally, we provide DEMO, a strong baseline model along with comprehensive experiments, demonstrating the effectiveness of our approach and fostering future research in this area.

2. Related Work

3D Human Motion-Language Datasets. Recent years have seen the emergence of datasets designed to advance research in 3D human motion generation and understanding, particularly those that pair motion data with natural language descriptions, with the first effort being the KIT-ML dataset [30]. Subsequent efforts [10, 31], significantly scaled the scope of motion-language datasets through crowdsourced annotation of 3D motion clips derived from existing mocap sequences, including AMASS [22] and HumanAct12 [9]. BABEL [31] annotates motion clips at two abstraction levels: overall sequence categories (*e.g.*, "play basketball") and subsequence action labels accompanied by durations (*e.g.*, "dribble ball with left hand", "run"), while many of which contain "transition" in-between. In HumanML3D [10], each motion clip is instead treated as a single semantic unit and described with three natural language sentences from different annotators. In contrast, the recent FineMotion [42] re-annotates the same motion sequences in HumanML3D, but segmenting them at uniform temporal intervals, irrespective of action semantics. Each snippet is labeled with fine-grained body-part movement descriptions (*e.g.*, "raise your hands up to your head") rather than action-centric labels or descriptions. MotionX [21] and its successor MotionX++ [51] shift the emphasis from more detailed captions toward enriching modalities. MotionX uses SMPL-X whole body pose annotations, covering body, hands, and facial expressions, paired with semantic labels. MotionX++ goes further by adding synchronized RGB video and audio data alongside pose annotations and textual descriptions. We propose CompMo, focusing on dense 3D human motion captioning. Rather than short labels like those in BABEL, coarse whole-clip captions like in HumanML3D, or snippet-level body-part descriptions as in FineMotion, our dataset provides rich sequence-level natural language descriptions, each annotated with precise temporal timestamps. CompMo thus establishes a new benchmark for dense motion captioning and motion-language alignment in 3D human motion, an area not yet addressed by existing datasets.

Dense Video Captioning. Dense Video Captioning (DVC) extends standard video captioning by identifying multiple temporal segments in an untrimmed video and generating corresponding textual descriptions for each segment [17]. Earlier methods typically followed a two-stage, detect-then-describe paradigm [17, 39], whereas recent approaches have shifted towards end-to-end training for improved efficiency and performance [5, 44, 48]. Effective DVC requires both accurate temporal localization and semantic correctness, and evaluation metrics must account for both aspects. To address this, DVC evaluation typically combines standard captioning metrics [2, 37] with Intersection over Union (IoU) thresholds. More recently, SODA [6] has been introduced as a comprehensive metric that temporally aligns predicted and reference captions before computing METEOR-based scores that penalize redundancy and poor alignment. We propose Dense Motion Captioning (DMC), bringing this paradigm to the domain of 3D human motion understanding, challenging models to generate temporally precise descriptions of human motion.

Human Motion Understanding. Much of prior work in human motion research has focused on motion generation [15, 26, 27, 32, 36], *i.e.*, synthesizing realistic 3D human movements from text or other modalities. More recently, the motion-to-text task has also gained attention, with methods developing unified motion-language models capable of both generating motion from text and describing input motion [4, 11, 15, 19, 35, 43, 47, 52]. While these demonstrate impressive versatility, their accuracy in motion understanding remains limited, particularly in tasks requiring temporal precision. This limitation arises because they are not trained to capture or describe sub-sequences within longer, continuous motions, which is essential for detailed temporal comprehension.

Beyond this, some works explore related but distinct challenges. BABEL-TAL [34] tackles 3D temporal action localization, which involves recognizing actions performed in a 3D motion sequence and precisely identifying their start and end

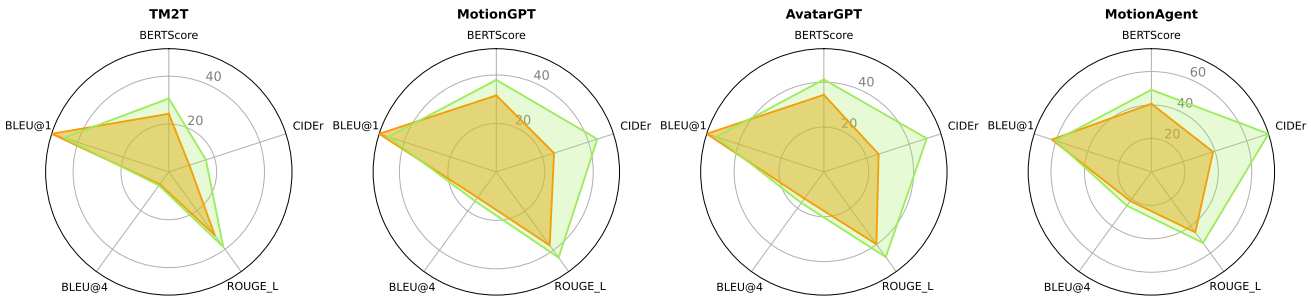

Figure 2. **Single Motion Captioning performance divided by *simple* and *complex* motion sequences.** We report the single motion captioning performance of state-of-the-art motion-language models on the simple and complex subsets of HumanML3D [10]'s test set, as defined in Sec. 3.1.

times, albeit with a fixed set of action class labels. Similarly, TMR [28] shows the use case of moment retrieval by temporally localizing BABEL actions within long sequences. This idea is later extended by UniMotion [18] to frame-level motion captioning as an initial exploration of dense action recognition. However, UniMotion [18] treats captioning as a retrieval problem with a closed vocabulary of action labels, and they do not provide a quantitative benchmark. In contrast, our method generates free-form descriptions and outputs segment timestamps instead of assigning an action label per frame.

3. From Simple to Complex Motions

In this section, we first motivate our study with a preliminary analysis of the widely used HumanML3D dataset [10] (Sec. 3.1). We then describe the generation pipeline of our dataset (Sec. 3.2).

3.1. Can Current Models Understand Complex Human Motions?

The HumanML3D [10] dataset is widely used to evaluate human motion understanding models thanks to its diverse range of motion sequences of varying complexity. In this study, we investigate whether the complexity of a motion, specifically, the presence of multiple sub-actions, correlates with the performance of state-of-the-art motion-language models. To this end, we partition the mirrored augmented dataset with 29,228 motions into two disjoint subsets: *simple* and *complex* motions. This partitioning is based on the number of verbs/adverbs in the ground-truth textual descriptions, under the assumption that each verb typically corresponds to a distinct sub-action (*e.g.*, "a person *sits* down and *crosses* their leg, before *getting up*"). Motions described with no more than 1 verb are considered simple, while those with 2 or more are labeled complex. This results in 17,512 complex and 11,716 simple motion instances, of which 2,663 and 1,721 are from the test set, respectively. We evaluate the performance of several recent models for standard single motion captioning, *i.e.*, generating one description without timestamps, [11, 16, 43, 52] on both subsets.[1] Fig. 2 reports the obtained results in terms of single motion captioning metrics [11]. In the vast majority of cases, we observe a considerable drop in performance on the complex subset, highlighting that current state-of-the-art models tend to perform better on simpler samples but struggle to accurately understand and describe longer sequences with multiple sub-actions. **This finding motivates our study**, emphasizing the need for datasets that present greater temporal complexity to better train and evaluate motion-language models, ultimately enabling more precise temporal motion understanding.

Dataset	Dataset Size	Avg. Duration (s)	Annotation Type	Timestamps
KIT-ML [30]	3,911	10.33	Sentence	✗
HumanML3D [10]	14,616	7.1	Sentence	✗
(mirror)	29,228			
BABEL [31]	13,220	12.26	Labels	✓
MotionX [21]	81,084	6.4	Sentence	✗
MotionX++ [51]	120,462	5.4	Sentence	✗
FineMotion [42]	14,616	7.1	Fine Descriptions	✓
CompMo (ours)	60,000	39.88	Dense Captions	✓

Table 1. **Overview of CompMo and prior 3D motion-language datasets.** While existing datasets vary in size, annotation type, and temporal richness, our CompMo is the first large-scale dataset designed for DMC with accurate timestamps, enabling more comprehensive modeling of temporally complex motions.

3.2. CompMo: A Complex Motion Dataset

To address the limitations current models face in handling temporally complex motions, we introduce the Complex Motion Dataset (CompMo), a new large-scale dataset specifically designed to challenge and advance motion-language models. CompMo is the first dataset explicitly created for 3D dense motion captioning with precise timestamps, enabling more effective training and evaluation of models. It features longer motion sequences, providing more temporally extended

[1] We exclude models that have not released code at the time of writing.

contexts for dense captioning. On average, each motion in CompMo is annotated with 37.74 words, compared to 12 and 11.06 words in HumanML3D [10] and BABEL [31]. Compared to existing temporally annotated motion datasets, CompMo represents a significant increase in both scale and complexity (see Tab. 1). To support these design goals, we developed a multi-stage pipeline for dataset construction, which we describe in detail below.

Atomic Actions Collection. To build a diverse and high-quality dataset for dense motion captioning, we begin by collecting simple human motions paired with textual descriptions. We use HumanML3D [10] as our primary source, as it provides an extensive collection of motion-text pairs encompassing a wide range of human motions, including everyday activities, sports, and artistic movements. Following our preliminary analysis (Sec. 3.1), we employ the *simple* set, treating each element as an atomic action aligned with its corresponding atomic description.

To obtain better alignment between motion and text, we propose two strategies for data collection: **i)** generated from scratch, and **ii)** drawn directly from the *simple* set. For the data in **i)**, we use the diffusion-based MDM-SMPL model proposed in STMC [29] to generate the motions from their textual descriptions; Then we use TMR [28], a model that encodes motions and languages into a shared embedding space, as encoder, to calculate the cosine *TMR Similarity* across different modalities, and filter out candidates with low motion-text alignment. To address motion types that are poorly generated, we supplement the dataset with samples from the *simple* HumanML3D set. The final atomic actions, accompanied by descriptions, contain 7,503 generated from scratch and 3,619 drawn from HumanML3D.

Textual Descriptions Composition. Starting from atomic actions, we perform a temporal composition for atomic descriptions by randomly sampling 2 to 10 atomics and combining these into coherent sequences. Each sequence is annotated with precise timestamps, formatted as "`<mm:ss:ms: atomic textual description>`". To ensure realistic and varied durations, we condition the length of each motion segment on its ground-truth duration from HumanML3D, applying small random perturbations to introduce variability while preserving temporal plausibility.

Motion Sequences Generation. We then generate human motion sequences corresponding to the constructed textual descriptions. Inspired by STMC [29], which applies a test-time denoising approach for spatio-temporal motion composition, we also employ the temporal stitching technique of DiffCollage [49] as well as the body part stitching in combination with MDM-SMPL provided by [29]. At each denoising step, we start from the textual description, denoise, stitch the resulting conditions together both temporally and across the relevant body parts, and finally generate the composed motion sequences.

Final Dataset Description. The resulting CompMo dataset contains 60,000 motion-text pairs with timestamp annotations. On average, motion sequences last 39.88 seconds, significantly longer than sequences in existing datasets, reflecting the increased temporal complexity of CompMo. We partition the dataset into training, validation, and test sets, corresponding to the 80%/10%/10% of the data, respectively. Additional details on the generation pipeline are provided in Sec. **??** of the Appendix.

4. DEMO: Dense Motion Captioning Model

In this section, we first formalize the dense motion captioning task (Sec. 4.1) then detail our proposed architecture (Sec. 4.2 and Sec. 4.3) and training procedure (Sec. 4.4).

4.1. Problem Formulation

Given a 3D human motion sequence $m \in \mathbb{R}^{N \times D}$, where N is the number of poses and D is the dimensionality of each pose, Dense Motion Captioning (DMC) consists in generating a sequence $\{(t_i, c_i)\}_{i=1}^{M}$, where $t_i = (s_i, e_i) \in \mathbb{R}^2$ represents the start and end times of the i-th motion segment, c_i is a caption describing the human motion within that segment, and M is the number of atomic actions detected. We define the pose dimensionality as $D = J \times 3$, where J is the number of 3D joints used to represent each pose. Unlike the traditional single motion captioning task, DMC requires both accurate temporal localization of atomic motion segments and natural language generation.

4.2. Method

Our architecture, DEMO, leverages an LLM, finetuned to autoregressively generate dense, temporally aligned captions from long and complex 3D motion sequences, as illustrated in Fig. 3 (left). Let f_ϕ denote the LLM, parametrized by ϕ. Since f_ϕ is originally pretrained only on text and vision modalities, it cannot directly process motion data. To address this, we first convert the continuous motion sequence $m \in \mathbb{R}^{N \times D}$ into a language-compatible embedding space that can be processed by f_ϕ, and then use f_ϕ to generate the dense motion descriptions.

4.3. Motion Representation

Prior LLM-based approaches represent a continuous motion by learning a mapping to discrete tokens, *e.g.*, training a vector quantized variational autoencoder (VQ-VAE) to construct a *motion vocabulary*. However, this approach suffers from two key limitations: (i) inherent information loss caused by the limited discrete vocabulary [23, 36], and (ii) the need for an additional, separate training stage for the VQ-VAE. In contrast, DEMO learns a simple continuous mapping from motion to language space using a single network. Specifically, a lightweight motion encoder γ extracts motion features, which are then adapted into the language domain via a linear projection $\mathbf{W}$, eliminating intermediate discretization.

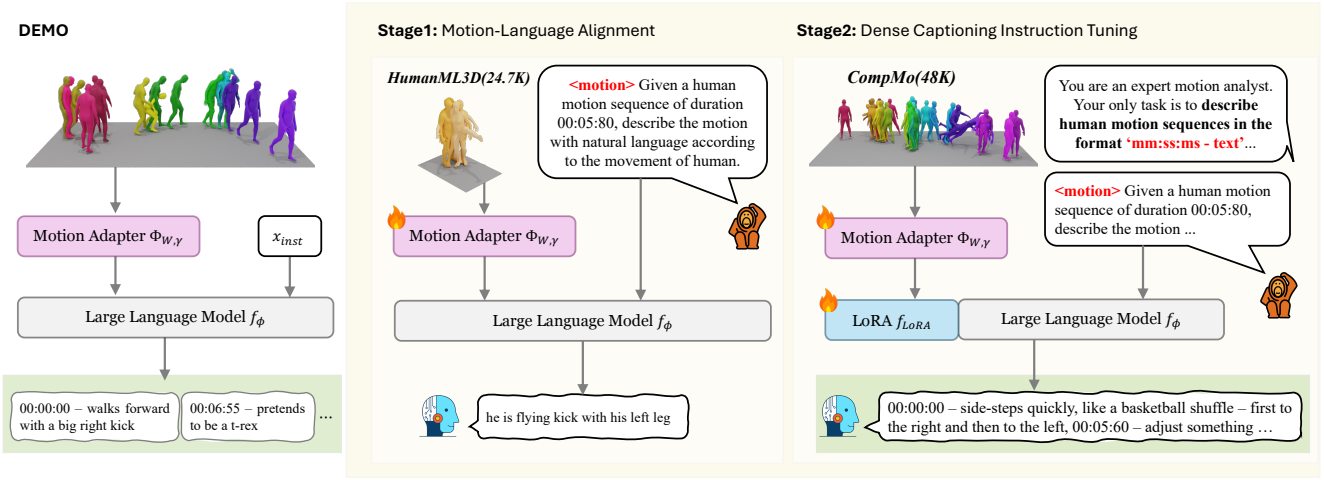

Figure 3. **DEMO overview :** Given a motion sequence m, our method encodes it with the motion adapter $\Phi_{\mathbf{W},\gamma}$, which maps it into the language embedding space of the LLM f_ϕ. Using the resulting motion embeddings and a textual instruction x_{inst}, the model generates dense captions with temporal boundaries. Training is conducted in two stages. Here, 🔥 denotes the subset of parameters being trained.

Since the motion sequences in CompMo can last up to 10 times the duration of those in HumanML3D, this necessitates a scalable and efficient strategy for encoding long sequences. Processing the entire motion at once is computationally expensive and often unnecessary, as generating detailed descriptions for short motion segments typically depends only on their immediate temporal context rather than the full sequence.

To address this, we partition the input motion sequence into a series of fixed-size, overlapping windows $\{m^{(i)} \in \mathbb{R}^{W\times D}\}_{i=1}^{K}$, extracted with a stride $S < W$. Each window $m^{(i)}$ corresponds to a sub-sequence of the full motion m and is processed independently to capture temporally localized motion patterns. The window is first flattened, added with positional embeddings, and then passed through the motion adapter defined as:

$$\Phi_{\gamma,\mathbf{W}}(m^{(i)}) = \mathbf{W}\cdot\gamma(m^{(i)}), \tag{1}$$

where the adapter projects the motion features into the language embedding space of f_ϕ.

4.4. Training Strategy

We train DEMO to autoregressively generate motion captions given a 3D motion sequence and a textual instruction. Given an input motion sequence m and instruction prompt x_{inst} as input, the generation process is modeled as:

$$p(\mathbf{y}\,|\,m,x_{inst}) = \prod_{i=1}^{L} p_\theta(y_i|m,x_{inst},y_{<i}), \tag{2}$$

where $\mathbf{y} = \{y_1,\ldots,y_L\}$ is the output caption of length L, $p(\cdot)$ is the model's probability distribution over tokens, and $y_{<i}$ denotes the previously generated tokens up to position $i-1$. The parameter set θ includes all trainable components of the model. During training, we optimize the model by maximizing the log-likelihood of the target caption, using the cross-entropy loss:

$$\mathcal{L} = -\sum_{i-1}^{L} \log p_\theta(y_i\,|\,m,x_{inst},y_{<i}). \tag{3}$$

As illustrated in Fig. 3, the motion adapter $\Phi_{\gamma,\mathbf{W}}$ and the LLM f_ϕ are trained in a two-stage process: first, a motion-language alignment stage to align motion features with the language model's embedding space, followed by a dense caption instruction tuning stage to enable precise and temporally grounded caption generation. While the training objective remains the same as in Eq. (3) in both stages, what differs are the instruction prompts x_{inst}, target outputs $\mathbf{y}$, input motion data m, and the subsets of parameters in θ optimized during training. These stages are described in detail below.

Stage 1: Pretraining for Motion-Language Alignment. In this stage, we focus on aligning the motion modality with the language space by training only the motion adapter, *i.e.*, $\theta = \Phi_{\gamma,\mathbf{W}}$ on paired motion-text data. To achieve this alignment, we use the HumanML3D [10], where each motion m consists of a single motion sequence, and $\mathbf{y}$ is the paired ground truth annotation, without timestamps. The instruction prompt x_{inst} is designed as shown in Fig. 3 (center), providing only the overall motion duration.

Stage 2: Dense Captioning Instruction Tuning. In this stage, we instruct the model to generate temporally grounded captions, explicitly including action boundaries and their corresponding timestamps. We use CompMo, where each motion m is a longer, complex sequence, and the target output $\mathbf{y}$ is a sequence of captions paired with their annotated temporal intervals. The instruction prompt x_{inst} is adapted

accordingly to guide the model in producing temporally localized descriptions, as illustrated in Fig. 3 (right). To enable efficient finetuning, we apply LoRA [14] to the language model f_ϕ, while jointly finetuning the pretrained motion adapter along with LoRA. Thus, the set of trainable parameters in this stage is $\theta = \{\Phi_\gamma, \mathbf{w}, f_{LoRA}\}$. This stage equips the model with the ability to generate fine-grained, time-aware descriptions of complex motions.

5. Experiments

Datasets and Settings. We conduct DMC experiments on two datasets: our proposed **CompMo**, and the intersection of HumanML3D [10] and BABEL [31], following the setup introduced in UniMotion [18]. CompMo comprises 60,000 motion sequences paired with dense captions, divided into 48,000/12,000 for training/testing. The dataset adopted from [18], here denoted with **H3D ∩ BABEL**, is constructed from the overlapping subset of HumanML3D and BABEL, and consists of 7,056/1,325 motion sequences paired with frame-level annotations for training/testing. Additionally, we use **HumanML3D** for the first stage of our training. We adopt the train+val split of the mirrored augmented dataset, including 23384+1460 motion sequences, each annotated with three descriptions. During training, we randomly sample one of the associated annotations at each step.

Metrics. We quantify DMC performance using dense captioning accuracy, temporal localization accuracy, and motion-caption alignment. For dense captioning, we follow dense video captioning literature [11, 17, 38, 44], computing captioning metrics: CIDEr [37], METEOR [2], ROUGE_L [20], BLEU [25], over matched prediction-reference pairs within the IoU thresholds of $\{0.3, 0.5, 0.7, 0.9\}$, reporting the average results on the matched pairs. We also use SODA [6] with two different linguistic metrics, METEOR [2] and BertScore [50] (corresponding to SODA and SODA(B) in Tab. 2), for overall caption evaluation. For temporal localization, we follow [38], using a greedy algorithm to select the best matching with the highest IoU, then computing the mean IoU for all matched pairs to get the overall tIoU and F1 score. For motion-caption alignment, following prior work on image and video captioning [17, 45], we measure the cross-modal distance between motion sequences and their generated captions. Specifically, we calculate the cosine similarity between motions and texts in the joint embedding space of TMR [28]. To further assess the sequential alignment, we adopt the CAR [7] score, a recent work that improves the motion-text retrieval by introducing negative samples generated through event-sequence shuffling, encouraging the model to achieve better temporal alignment, where we retrieve motions given a set of shuffled and generated event sequence captions from the test set with 32 samples.

Implementation Details. We use 3D joint representations with $J = 22$ joints. We set the window size and stride to $W = 16$, $S = 8$. Our f_ϕ is initialized with LLaMA-3.1-8B-Instruct [8], while γ is an MLP. Training takes approximately 3.5 hours on 2 NVIDIA RTX 6000 Ada GPUs. Additional implementation details are provided in Sec. **??** of the Appendix.

5.1. Comparative Results

Quantitative Results. To the best of our knowledge, dense motion captioning is a novel task that has not been systematically addressed and evaluated in prior work. For comparison, we adapt UniMotion [18] as a baseline for our evaluations. While UniMotion does not produce dense captions, we aggregate its frame-level predictions into temporal segments for fair comparison. Tab. 2 reports quantitative results for both our proposed DEMO and UniMotion trained and tested on the CompMo and H3D ∩ BABEL datasets, where UniMotion previously provided only qualitative examples. DEMO outperforms UniMotion, particularly on the more challenging CompMo. It achieves better temporal localization performance on both datasets, with $+34.1/3.9\%$ improvements in tIoU, and shows substantial gains in dense captioning quality, *i.e.*, $+13.2/5.1\%$ on SODA metrics. This performance gap can be attributed to fundamental differences in methodology: UniMotion predicts CLIP embeddings for frame-level text descriptions and retrieves captions from a pre-computed vocabulary using a K-nearest neighbor search. This pipeline requires prior knowledge of the dataset's action labels. Moreover, when the vocabulary of potential action descriptions is large (CompMo contains 11,085 atomic actions compared to 6,133 in H3D ∩ BABEL), this approach is limited by the effectiveness of the retrieval process. Additionally, because UniMotion relies on CLIP, it is subject to CLIP's token limit of 77 tokens per text input [24]. This limitation truncates longer descriptions, significantly hindering performance on more detailed captions. In contrast, DEMO directly generates captions in an open-ended manner, avoiding these constraints. As a result, on CompMo, which features longer and more semantically rich descriptions compared to H3D ∩ BABEL, DEMO outperforms UniMotion, particularly on dense captioning metrics.

Qualitative Results. Fig. 4 presents a qualitative comparison between DEMO and UniMotion on the challenging CompMo. The results indicate that DEMO generates more accurate segments of action boundaries and produces captions that align better with the ground-truth annotations in style. For example, it often divides motion sequences into the correct number of atomic actions, with only occasional omissions (*e.g.*, missing one step in the top example). Furthermore, it accurately captions the depicted actions in most instances, while UniMotion's frame-level captions often contain noise and fail to accurately describe the actions. Interestingly, in some cases, the generated descriptions differ from the ground truth in wording but still convey an equivalent meaning (*e.g.*, generating "*kicks with their right leg four times while their hands are in front of their face*" instead of "*doing karate kicks*" in the bottom example). More results can be found

Method	Dataset	Dense Captioning ↑							Localization ↑		T-M Similarity ↑	
		SODA	SODA(B)	CIDEr	METEOR	ROUGE_L	BLEU@1	BLEU@4	tIoU %	F1 %	TMR	CAR
UniMotion [18]	CompMo	0.6099	12.8090	1.0082	0.4266	0.8479	0.7793	0.0000	36.14	4.00	0.4930	0.3487
DEMO	CompMo	17.8473	64.4003	134.4424	16.4085	24.0469	23.8980	11.0024	77.94	58.21	0.6832	0.8027
UniMotion [18]	H3D ∩ BABEL	5.7141	30.4658	6.7170	5.0826	5.8060	5.1651	0.4375	49.95	22.23	0.6428	0.8473
DEMO	H3D ∩ BABEL	7.9194	25.9654	7.8090	5.7625	6.2919	5.6936	0.1318	51.56	16.40	0.6052	0.8204

Table 2. **Comparison on Dense Motion Captioning**. We compare the performance of DEMO on the proposed CompMo and on H3D ∩ BABEL. We measure dense captioning, temporal localization, and motion-caption alignment accuracy. Best results are highlighted.

Method	Dense Captioning ↑							Localization ↑		T-M Similarity ↑	
	SODA	SODA(B)	CIDEr	METEOR	ROUGE_L	BLEU@1	BLEU@4	tIoU %	F1 %	TMR	CAR
Dataset Generation											
Concat GT	1.9910	41.5498	8.2427	1.9572	4.0158	4.2401	0.0428	61.45	27.52	0.5414	0.4505
Smooth GT	1.9561	41.5586	8.1089	1.8835	3.9398	4.1223	0.0230	61.08	26.74	0.5306	0.4977
Denoise only from random	12.1643	62.4457	80.9095	11.9174	18.2653	18.3024	5.1632	77.92	57.32	0.5680	0.7895
Denoise only from GT	13.3860	55.2276	94.7040	12.7457	17.5265	17.7187	7.6551	69.89	43.00	0.5754	0.7987
CompMo	17.8473	64.4003	134.4424	16.4085	24.0469	23.8980	11.0024	77.94	58.21	0.6832	0.8027
Training Stages											
Stage 2	1.6521	28.4648	4.5059	1.2444	2.0972	2.3754	0.0362	49.45	14.28	0.6056	0.5987
Stage 1+2	17.8473	64.4003	134.4424	16.4085	24.0469	23.8980	11.0024	77.94	58.21	0.6832	0.8027
Motion Representation											
VQ-VAE	2.3398	43.3563	7.6868	2.0440	3.4973	3.6243	0.0778	60.76	26.60	0.5881	0.6282
$\Phi_{\mathbf{W},\gamma}$	17.8473	64.4003	134.4424	16.4085	24.0469	23.8980	11.0024	77.94	58.21	0.6832	0.8027

Table 3. **Ablation Study.** We assess the contribution of different components by ablating variations in dataset generation (data-level), as well as training stages and motion representation (model-level). The grey-highlighted configuration corresponds to the one used in our final model and full data pipeline.

in the provided supplementary video.

5.2. Ablation Study

In this section, we examine the key factors that influence the DMC performance. We first study the impact of our dataset generation strategy, followed by an evaluation of our training strategy. Finally, we investigate how different motion representations affect the results. Additional details are provided in Sec. **??** of the Appendix.

Dataset Generation. To evaluate the effectiveness of our proposed data generation strategy, we ablate different components of the pipeline and train our DEMO on the resulting variant datasets. To evaluate the role of atomic actions collection strategies in Sec. 3.2, we compare two modes: (i) solely generated from scratch (*denoise only from random*); and (ii) solely drawn from HumanML3D (*denoise only from GT*); then resample and denoise for sequences composition based on these two atomic actions. To examine the role of denoising in generating and composing motion sequences, we also create data by directly concatenating HumanML3D motions without denoising (*concat GT*), and apply a smoothed version using Slerp interpolation [33] (*smooth GT*). As shown in Tab. 3, the proposed mixture-denoising strategy consistently yields superior performance, demonstrating that it produces higher-quality datasets for training the DMC model.

Training Strategy. To assess the impact of our proposed two-stage training strategy in Sec. 4.4, we ablate the motion-language alignment stage and finetune the LLM directly on CompMo (*stage 2 only*). In this setting, the LLM is adapted with LoRA, while the motion adapter is randomly initialized and trained from scratch together with the LLM. As we reported in Tab. 3, the full pipeline (*stage 1+2*)significantly improves the results in both temporal localization (+20.8% tIoU) and dense captioning accuracy (+12.1% SODA), underscoring the importance of motion-language alignment prior to LLM finetuning.

Motion Representation. Prior LLM-based methods [11, 16, 40, 43] adopt VQ-VAE to discretize motion into token sequences, which introduces an additional training stage and restricts input motions to short sequences (*i.e.*, up to 200 poses). Building on our prior discussion of motion representation in Sec. 4.3, we conduct an ablation study comparing our continuous motion encoding ($\Phi_{\mathbf{W},\gamma}$) to the conventional VQ-VAE tokenizer (*VQ-VAE*). For this experiment, we substitute our motion adapter $\Phi_{\mathbf{W},\gamma}$ with a VQ-VAE pretrained on HumanML3D [10]. This approach encodes motions into discrete token indices, which are then mapped back to their corresponding continuous feature vectors from the VQ-VAE's codebook before being passed to the LLM for further processing. We then train the model through the subsequent two stages: motion-language alignment on HumanML3D, followed by dense-caption instruction tuning on CompMo. The results in Tab. 3 show that the VQ-VAE-based model significantly underperforms ours, particularly on captioning metrics, highlighting the challenges posed by its limited discrete vocabulary in capturing the complexity of CompMo.

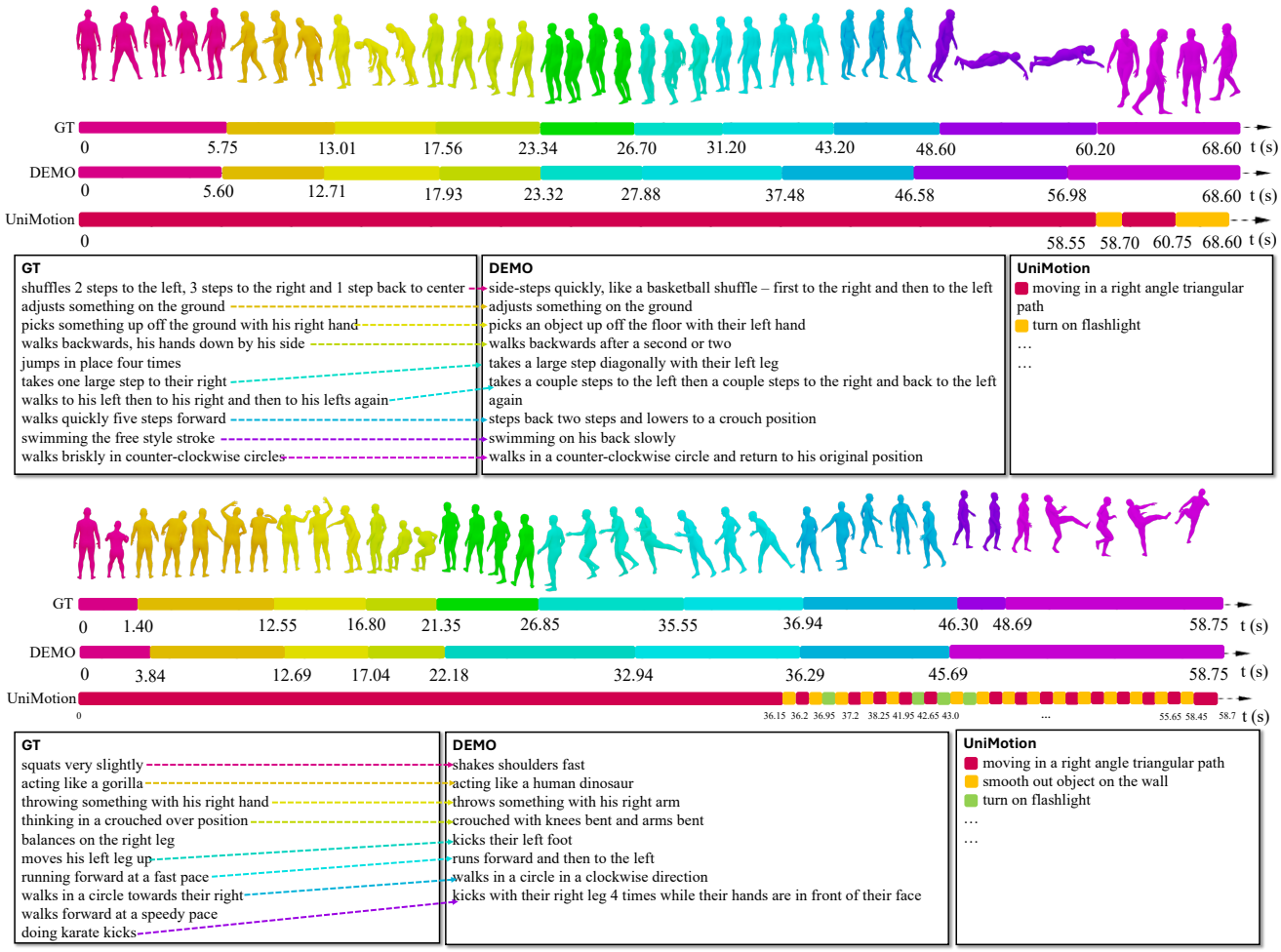

Figure 4. **Qualitative Results.** We show two motion sequence examples from the CompMo dataset, along with the ground truth annotations (GT) and the dense captions predicted by our DEMO and UniMotion. For each sequence, the top rows show the temporal intervals of the input motion divided according to the GT and the two model predictions, with the corresponding captions listed below. Predicted captions that align with the GT are highlighted in the same color and connected with arrows to indicate the alignment.

6. Conclusion

In this work, we propose the novel task of dense motion captioning, broadening the scope of 3D human motion understanding. To address the scarcity of suitable datasets for this task, we further introduce CompMo, a large-scale dataset of 3D long human motion sequences, annotated with temporal sequences of actions and timestamps. By enabling models to generate detailed motion descriptions from 3D data, this task supports the development of systems that can better understand human movement, *e.g.*, moving beyond raw RGB video analysis to a more precise understanding of motion itself, by lifting 2D videos into 3D human motion representations and interpreting the underlying actions.

While CompMo currently focuses on temporal composition of movements, future work could extend this to spatio-temporal composition and understanding. Moreover, it does not enforce any constraints on the temporal arrangement of actions, enabling the generation of random sequences. However, this can lead to incoherent compositions, for example abruptly switching from *swimming* to *playing basketball* without a plausible transition, since modeling causal relationships between actions is outside the scope of this work. A promising direction for further dataset improvements is to incorporate realistic long-term behaviors, such as multiple sub-actions related to basketball or other complex, structured human motions. This could enable models to caption motion sequences that more faithfully emulate natural human movement.

Acknowledgments. The authors acknowledge the ANR project CorVis ANR-21-CE23-0003-01 CorVis. The authors thank the Deep Learning Lab of the ProM Facility for the GPU time. This work is also supported by the EU Horizon project ELLIOT (No. 10121439).

References

[1] Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3d human motion editing. In *SIGGRAPH Asia*, 2024. 1

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005. 2, 6

[3] Léore Bensabath, Mathis Petrovich, and Gül Varol. TMR++: A cross-dataset study for text-based 3d human motion retrieval. In *CVPRW HuMoGen*, 2024. 1

[4] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. MotionLLM: Understanding human behaviors from human motions and videos. *arXiv:2405.20340*, 2024. 2

[5] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *CVPR*, 2021. 2

[6] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation framework. In *ECCV*, 2020. 2, 6

[7] Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models. In *ECCV*, 2024. 1, 6

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 6

[9] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACMMM*, 2020. 2

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7

[11] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In *ECCV*, 2022. 1, 2, 3, 6, 7

[12] Ziyan Guo, Zeyu Hu, Na Zhao, and De Wen Soh. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *ICCV*, 2025. 1

[13] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. SALAD: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *CVPR*, 2025. 1

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 6

[15] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. CoMo: Controllable motion generation through language guided pose code editing. In *ECCV*, 2024. 2

[16] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. *NeurIPS*, 2024. 1, 3, 7

[17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 2, 6

[18] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3D human motion synthesis and understanding. In *3DV*, 2025. 3, 6, 7

[19] Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, and Jenq-Neng Hwang. Human motion instruction tuning. *arXiv:2411.16805*, 2024. 2

[20] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 6

[21] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A large-scale 3D expressive whole-body human motion dataset. *NeurIPS*, 2023. 1, 2, 3

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2

[23] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. In *CVPR*, 2025. 4

[24] Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M Asano, Nanne Van Noord, Marcel Worring, and Cees GM Snoek. Tulip: Token-length upgraded clip. *arXiv:2410.10034*, 2024. 6

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[26] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer vae. In *ICCV*, 2021. 2

[27] Mathis Petrovich, Michael J Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 1, 2

[28] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 1, 3, 4, 6

[29] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR Workshop on Human Motion Generation*, 2024. 4

[30] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 2016. 1, 2, 3

[31] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *CVPR*, 2021. 1, 2, 3, 4, 6

[32] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024. 2

[33] Ken Shoemake. Animating rotation with quaternion curves. 19(3), 1985. 7

[34] Jiankai Sun, Linjiang Huang, Jianing Qiu Hongsong Wang, Chuanyang Zheng, Md Tauhidul Islam, Enze Xie, Bolei Zhou, Lei Xing, Arjun Chandrasekaran, and Michael J. Black. Localization and recognition of human action in 3D using transformers. *Nature Communications Engineering*, 2024. 2

[35] Shanlin Sun, Gabriel De Araujo, Jiaqi Xu, Shenghan Zhou, Hanwen Zhang, Ziheng Huang, Chenyu You, and Xiaohui

Xie. CoMA: Compositional human motion generation with multi-modal agents. *arXiv:2412.07320*, 2024. 2

[36] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 4

[37] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 6

[38] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Chapter-Llama: Efficient chaptering in hour-long videos with LLMs. In *CVPR*, 2025. 6

[39] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018. 2

[40] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. MotionGPT-2: A general-purpose motion-language model for motion generation and understanding. *arXiv:2410.21747*, 2024. 7

[41] Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick W. B. Li, Ziyao Zhang, and Xiaohui Liang. Fg-t2m++: Llms-augmented fine-grained text driven human motion generation. *Int. J. Comput. Vision*, 2025. 1

[42] Bizhu Wu, Jinheng Xie, Meidan Ding, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. FineMotion: A dataset and benchmark with both spatial and temporal annotation for fine-grained motion generation and editing. *arXiv:2507.19850*, 2025. 2, 3

[43] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with LLMs. In *ICLR*, 2025. 1, 2, 3, 7

[44] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 2, 6

[45] Hongkuan Zhang, Saku Sugawara, Akiko Aizawa, Lei Zhou, Ryohei Sasano, and Koichi Takeda. Cross-modal similarity-based curriculum learning for image captioning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7599–7606, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 6

[46] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 1

[47] Pengfei Zhang, Pinxin Liu, Hyeongwoo Kim, Pablo Garrido, and Bindita Chaudhuri. KinMo: Kinematic-aware human motion understanding and generation. *arXiv:2411.15472*, 2024. 2

[48] Qi Zhang, Yuqing Song, and Qin Jin. Unifying event detection and captioning as sequence generation via pre-training. In *ECCV*, 2022. 2

[49] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming yu Liu. DiffCollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023. 4

[50] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020. 6

[51] Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X++: A large-scale multimodal 3D whole-body human motion dataset. *arXiv:2501.05098*, 2025. 2, 3

[52] Zixiang Zhou, Yu Wan, and Baoyuan Wang. AvatarGPT: All-in-one framework for motion understanding, planning, generation and beyond. In *CVPR*, 2024. 1, 2, 3