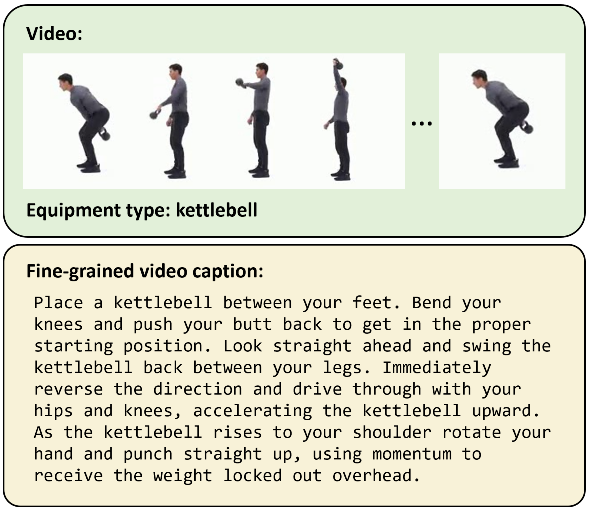
Exploring Fine-Grained Human Motion Video Captioning

Bingchan Zhao[1*], **Xinyi Liu**[1*], **Zhuocheng Yu**[1], **Tongchen Yang**[1],
Yifan Song[1], **Mingyu Jin**[1], **Sujian Li**[1,2†], **Yizhou Wang**[1]
[1] National Key Laboratory for Multimedia Information Processing, Peking University
[2]Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University
{bingchan.zhao, lisujian}@pku.edu.cn, liuxy0406@stu.pku.edu.cn

Abstract

Detailed descriptions of human motion are crucial for effective fitness training, which highlights the importance of research in fine-grained human motion video captioning. Existing video captioning models often fail to capture the nuanced semantics of videos, resulting in the generated descriptions that are coarse and lack details, especially when depicting human motions. To benchmark the **Bo**dy **Fi**tness **T**raining scenario, in this paper, we construct a fine-grained human motion video captioning dataset named BoFiT and design a state-of-the-art baseline model named BoFiTGen (**Bo**dy **Fi**tness Training **T**ext **Gen**eration). BoFiTGen makes use of computer vision techniques to extract angular representations of human motions from videos and LLMs to generate fine-grained descriptions of human motions via prompting. Results show that BoFiTGen outperforms previous methods on comprehensive metrics. We aim for this dataset to serve as a useful evaluation set for visio-linguistic models and drive further progress in this field. Our dataset is released at https://github.com/colmon46/bofit.



Place a kettlebell between your feet. Bend your knees and push your butt back to get in the proper starting position. Look straight ahead and swing the kettlebell back between your legs. Immediately reverse the direction and drive through with your hips and knees, accelerating the kettlebell upward. As the kettlebell rises to your shoulder rotate your hand and punch straight up, using momentum to receive the weight locked out overhead.

Figure 1: One example in our dataset BoFiT. In previous work, only a one-sentence caption such as "A man demonstrates how to do a single arm snatch" is provided for the video.

1 Introduction

In today's fast-paced, high-stress lifestyles, many people aim to stay fit and healthy through self-training, either at the gym or at home. To get detailed guidance, they often rely on video courses. As we know, nuanced human motion descriptions from exercise videos are helpful to follow these videos and assess motion quality during training. However, providing detailed guidance for each body motion in videos can be costly and time-consuming. To address this challenge, research in fine-grained video captioning has become necessary, as it can automatically generate accurate, detailed descriptions of human motions, making high-quality fitness guidance more accessible.

Existing video captioning datasets involving human motions are mainly used in action recognition tasks, where each video is classified into a specific category (Kuehne et al., 2011; Soomro et al., 2012; Kay et al., 2017; Carreira et al., 2018, 2019; Smaira et al., 2020). Such video captioning can be seen as operating at the keyword level, which falls short of providing the fine-grained human motion descriptions needed for instructional purposes, such as detailed, step-by-step body motion analysis. Subsequently, several sports-specific video captioning datasets have been constructed, covering domains such as basketball, volleyball, and football competitions (Yu et al., 2018; Pasunuru and Bansal, 2018; Qi et al., 2019; Suglia et al., 2022). However, these datasets primarily focus on human interactions and do not address the fine-grained motions of body trunks.

Thus, in this paper, we propose a novel task of

[*] Equal Contribution.
[†] Sujian Li is the corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5247–5264
January 19–24, 2025. ©2025 Association for Computational Linguistics

fine-grained human motion video captioning to fill in the blanks of previous works. To benchmark the task, a video captioning dataset is necessary which should ensure the qualities of the videos and their corresponding fine-grained captions. On one hand, the fitness training videos must be professional, with high-caliber trainers. On the other hand, the motion videos should be accompanied by expert descriptions for guiding each body movement, though this involves a high workload and may be influenced by human subjectivity. With the considerations above, we build a dataset named BoFiT (Body Fitness Training Dataset), sourced from a professional fitness training website BodyBuilding[1] since it has clear and professional training videos some of which are equipped with fine-grained, body-trunk level descriptions. The main issue is that not all videos come with annotated descriptions. To address this, we use large language models (LLMs) combined with manual proofreading to generate fine-grained descriptions for those videos lacking them.

Now, BoFiT can serve as a testbed for evaluating the performance of fine-grained descriptions approaches. Our preliminary experiments show, previous VLMs (Vision-Language Model) (Luo et al., 2020; Lin et al., 2021; Tang et al., 2021; Seo et al., 2022; Li et al., 2022; Ye et al., 2022; Yan et al., 2022; Wang et al., 2022; **?**) and models enhanced with LLMs (Maaz et al., 2023; Zhang et al., 2023; Lin et al., 2023) still underperform on BoFiT by providing wrong depictions of human motion. To solve fine-grained video comprehension issues and fully leverage LLM's potentials on text generation, we develop an intermediate representation of human motion. This is achieved by extracting information from videos using a State-Of-The-Art 3D pose estimation model. This approach converts videos into systematic semantic representation, while also providing interpretable inputs for LLMs. Given LLMs' proven proficiency in reasoning over structured data such as tables and graphs (Hegselmann et al., 2023; Chen et al., 2024; Jin et al., 2024), it is plausible for us to formalize angular representation as a substitute for human motion in videos. We name this method of extracting human pose representations for LLM analysis as BoFiTGen. Based on BoFiT, we conduct in-depth experiments to investigate the performance of BoFiTGen and other video captioning models on different aspects. Results show that BoFiTGen outperforms others in comprehensive metrics.

Our contribution can be summarized as follows:

- We propose a novel fine-grained human motion video captioning task and correspondingly construct a semi-automatically labeled dataset BoFiT, which contains fitness training videos and their fine-grained descriptions at the body-trunk level.

- To address complex video captioning challenges, we propose the usage of human posture features as intermediate representations between video and text, helping LLMs understand videos more effectively.

- We design a few-shot LLM-based video captioning method called BoFiTGen, which successfully generates fine-grained instructional descriptions given fitness training videos. Results demonstrate the superior capability of BoFiTGen on the video captioning task.

2 Related Work

2.1 Fine-Grained Video Captioning

The task of dense video captioning is introduced by Krishna et al. (2017). It divides the untrimmed video into clips with the start and end frame, and attached captions related to a set of temporally localized activities. Among the existing dense video captioning tasks, those focusing on the sports domain are the most relative to our research focus. On one hand, some existing works formalize dense video captioning as (Krishna et al., 2017) does, aiming to generate short captions for trimmed video clips. The overall video would then be paired with aggregated dense captions as a whole. For example, Qi et al. (2019); Suglia et al. (2022) are benchmarks that pair trimmed football comment videos to captions with a length of one to two sentences. On the other hand, some works generate a fine-grained long caption for the entire video at once (Yu et al., 2018; Qi et al., 2019). They are closer to our research goal but fail to focus on describing body-trunk-level human motions, generating action-level sports descriptions instead. Here, we delve deeper into the granularity of human body trunks by constructing BoFiT as a more challenging task than before.

[1] https://www.bodybuilding.com

2.2 LLMs for Multi-modal Tasks

Recently, many works intend to extend LLMs to understand visual inputs including images and videos. Main approaches fall into two categories. The first category is to use LLMs as an agent to schedule and employ off-the-shelf expert models, such as captioning, retrieval, and OCR models (Shen et al., 2023; Wu et al., 2023; Surís et al., 2023; Yang et al., 2023). The second category is to use LLMs as a decoder. Fundamental large-scale vision-language models (VLMs) usually consist of a vision encoder, an LLM as a decoder, and a cross-modal interaction module to achieve vision-language alignment. For example, Flamingo (Alayrac et al., 2022) uses perceiver resampler and gated-cross attention while BLIP-2 (Li et al., 2023) uses Q-Former to adapt visual features for LLM. Subsequently, Dai et al. (2023); Liu et al. (2023); Zhu et al. (2023a) explore methods for visual instruction tuning and making VLMs more instruction-aware. Meanwhile, Zhang et al. (2023); Maaz et al. (2023); Lin et al. (2023) extend inputs from images to videos.

2.3 3D Human Pose Estimation

3D Human Pose Estimation focuses on extracting three-dimensional human poses from monocular RGB videos. Early works primarily explore single-stage approaches, which directly extract 3D pose information from input images (Sun et al., 2017; Moon et al., 2019; Zhou et al., 2019). More recently, two-stage methods have become widely adopted (Pavllo et al., 2019; Zhao et al., 2019; Zheng et al., 2021; Zhu et al., 2023b; Zhao et al., 2023) These methods initially use 2D pose estimators (Simonyan and Zisserman, 2014; He et al., 2015; Newell et al., 2016; Pang et al., 2018, 2020), followed by lifting these predictions to 3D poses. Notably, (Zhao et al., 2019) and (Pavllo et al., 2019) integrate semantic graphs and dilated temporal information into convolution networks to improve the estimation process. Moreover, Transformer (Vaswani, 2017) is also used to learn versatile temporal features from video sequences (Zheng et al., 2021; Zhu et al., 2023b; Zhao et al., 2023).

3 Task and Dataset Description

3.1 Fine-grained Video Captioning Task

Different from previous video captioning tasks in the sports domain, we propose a video captioning task which focuses on body-trunk-level human motion. Given a video clip V_i capturing the movement of an individual, our model is expected to generate a fine-grained description I_i of the motion, including the direction of movement for limbs and the final position reached. Figure 1 demonstrates a fitness training video with sequential human motions and our corresponding fine-grained target caption. Different from previous short captions, our BoFiT-Gen generates long captions that depict detailed human motion. To accompany the proposed task, we construct a dataset named BoFiT.

Dataset	**Scenario**	**Sentences per sec**	**Words per sec**
MSR-VTT	Open Domain	0.067	0.621
ActivityNet	Open Domain	0.327	4.410
YouCook2	Cooking	0.051	0.449
FSN	Basketball	0.556	4.901
SVCDV	Volleyball	0.366	3.886
BoFiT	Fitness Training	**2.819**	**48.224**

Table 1: Comparisons among video captioning datasets.

3.2 BoFiT Dataset

We collect 2360 videos from BodyBuiding, a professional fitness training instructional website. These videos have been provided with professional information including motion names, types, equipment, benefits, short descriptions, detailed instructions, etc. To minimize the estimation bias introduced by the vision model, we select those videos featuring a single person exercising without frequent switching from one shot to another. We carefully trim each video to contain only one cycle of training movement, as the original video may contain several ones.

We first considered getting annotated instructions from the BodyBuilding website to equip each video with a fine-grained caption. These instructions are of high quality, including detailed descriptions and tips for every step of the training movement. However, only 920 videos have such annotated and professional instructions among all 2360 items. For the 1440 videos without text instructions, it is difficult to manually compile professional instructions without expertise in the sports field. To promote the efficacy of instruction editing, we make use of the strong generation ability of ChatGPT and prompt it to generate instructions. During the ChatGPT-driven instruction generation process, we provide the motion name in the corresponding video and an expected instruction length in the prompt, which is set as the average length of the existing annotated instructions. This will en-

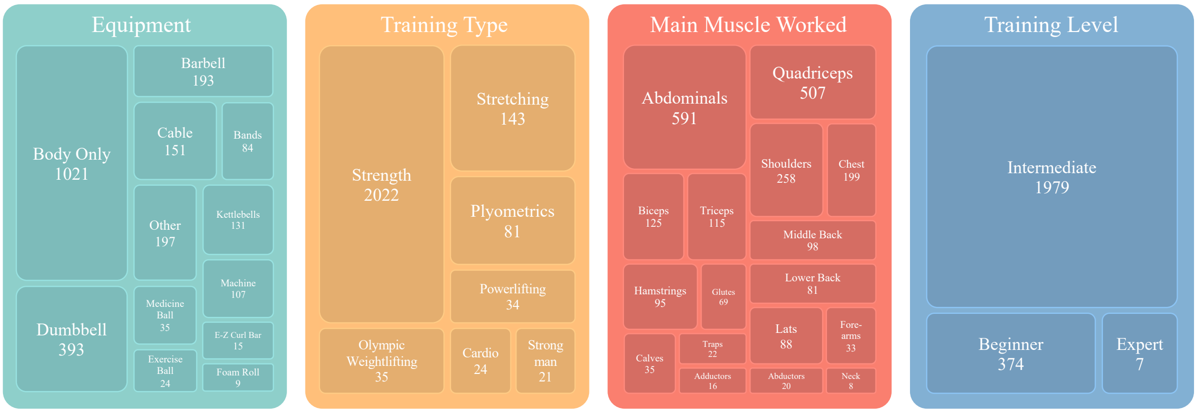


Figure 2: All properties featured in BoFiT. The video clips include different equipment and training types, mainly worked muscle and training levels.

sure that all generated instructions are independent of the visual content of the videos, sourced only from the motion name.

To ensure the data consistency between generated and annotated instructions, we build a website based on *Label Studio* to collect human feedback on them. For data consistency evaluation, we generate extra instructions for the 920 videos that already have expert-annotated instructions. For each video clip, we have an annotated instruction and a generated instruction. We randomly sample 200 instruction pairs and recruit 20 people to evaluate each pair twice. On the website, we ask the human judges to check if the two instructions are consistent, generally consistent, or inconsistent in meaning. Results show that 84.5% pairs are considered consistent or generally consistent by human judges. Therefore, we believe that our LLM-driven measure is qualified to help construct BoFiT dataset.

3.3 Dataset Statistics

BoFiT has 2360 video clips with a resolution of 480 × 270 at 30 fps . Each video clip spans 2.94 seconds on average and is paired with 8.3 sentences and 141.8 words on average. The comparison of BoFiT with other video captioning datasets (Heilbron et al., 2015; Xu et al., 2016; Zhou et al., 2017; Yu et al., 2018; Qi et al., 2019) is shown in Table 1. To the best of our knowledge, BoFiT provides the most abundant sentences and words per second among all datasets in both the open domain and sports domain.

The BoFiT dataset shows great diversity in its properties, including different equipment, training type, main muscle worked and training level. Detailed dataset composition is shown in Figure 2. In the collected 2,360 video clips, body-only training movement accounts for the largest percentage, up to 1021 videos. Other equipment includes dumbbell, barbell, and others. With regards to the training type of the movements, they mostly fall into the domain of strength training. Additionally, it includes stretching, plyometrics, powerlifting, Olympic weightlifting, cardio and strongman exercises. Furthermore, almost all muscles in the human body are trained in our BoFiT dataset. Finally, the training level of the videos are categorized into beginner, intermediate and expert levels, with intermediate-level videos taking up the largest proportion.

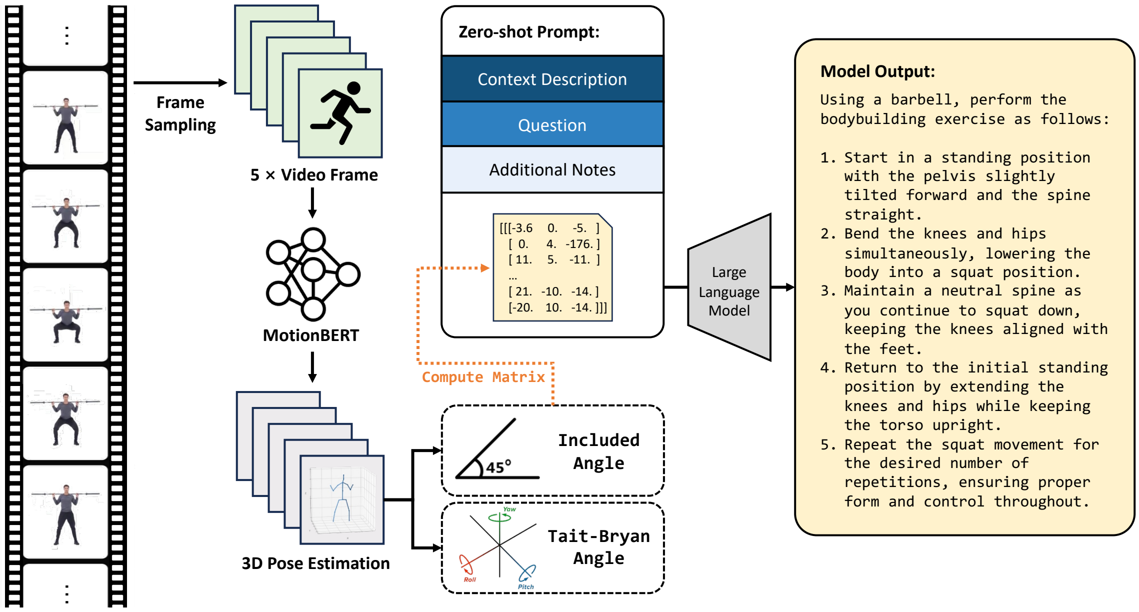
We believe our BoFiT dataset is effective in representing diverse human motions by engaging all body-trunks in the movements. Therefore, we believe methods built upon BoFiT dataset will have the potential to generalize on other human motion video captioning datasets and can also be developed for further practical use.

4 Method

We develop an LLM-based pipeline named BoFiT-Gen to address the challenge of fine-grained human motion video captioning. As demonstrated in Figure 3, it first extracts the angular data of the human motion in a given video clip through a State-Of-The-Art 3D human pose estimation model. It then encodes the data into a carefully designed prompt to generate fine-grained text descriptions through an LLM.

4.1 3D Human Pose Estimation

Here we employ MotionBERT (Zhu et al., 2023b) as the methodology for extracting 3D human mo-



Model Output:

```
Using a barbell, perform the
bodybuilding exercise as follows:

1. Start in a standing position
   with the pelvis slightly
   tilted forward and the spine
   straight.
2. Bend the knees and hips
   simultaneously, lowering the
   body into a squat position.
3. Maintain a neutral spine as
   you continue to squat down,
   keeping the knees aligned with
   the feet.
4. Return to the initial standing
   position by extending the
   knees and hips while keeping
   the torso upright.
5. Repeat the squat movement for
   the desired number of
   repetitions, ensuring proper
   form and control throughout.
```

Figure 3: An overview of BoFiTGen. The video clip in BoFiT is first sampled to 5 frames and then processed by MotionBERT to access the 3D pose estimation information. Extracted information included angles and Tait-Bryan angles, which are then computed into angular representation matrix and then used to generate fine-grained description via prompting Large Language Models.

tion information from the fitness training videos. This model can perform two key functions: On one hand, it can regress the 3D coordinates of the human joints at each frame. On the other hand, it can predict the local rotations of joints around its predecessors on the kinematic tree. Both the 3D coordinates and local rotations of the human joints are obtained for later use.

4.2 Included Angle Representation

We propose an angular representation system named Included Angle Representation that directly computes the angles between different pairs of limbs, with an assumption that the human skeleton is a composition of rigid bodies. We define a human coordinate system. The direction from the right hip to the left hip is denoted as the Y-axis, the direction from the midpoint of the pelvis to the lumbar vertebrae is the Z-axis, and the direction perpendicular to them is the X-axis.

All movements within this coordinate system are defined as local human motion. We classify the human joints into two categories according to their degrees of freedom (Akhter and Black, 2015) and employ different representation methods respectively. Afterwards, we define global human motion as a composition of jumping, rotating, and translating. The definition details of the included angle representation is demonstrated in Appendix.A. For each video V_i with N sampled frames, we obtain an overall included angle representation matrix $R_i \in \mathbb{R}^{N\times 22\times 3}$.

4.3 Tait-Bryan Angle Representation

In this section, we conduct another modeling system called Tait-Bryan Angle Representation. As per the definition of Euler Angles, we take a rotation in the 3D coordinate system as a sequence of three elementary rotations. In particular, Tait-Bryan Angles are sequential rotations made around three distinct rotation axis. Under our definition, axis $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ are of the body frame.

With the assistance of visual models, we obtain some local rotation quaternions predicted by the MLP regressor of the MotionBERT (Zhu et al., 2023b). These quaternions depict how each body joint rotates around its precedent on the kinematic tree. According to Berner et al. (2008), we transfer quaternions to Tait-Bryan angles. Equations are listed in Appendix.B. Additionally, we add a global information vector that includes the 3D coordinates of the pelvis (i.e. root node) in the global coordinate system. Finally, we obtain the Tait-Bryan representation matrix by concatenating the global and local rotation information at the feature dimension. For each video V_i with N sampled frames, we obtain an overall Tait-Bryan Angle Representation matrix $R_i \in \mathbb{R}^{N\times 17\times 3}$.

4.4 Fine-grained Text Generation via Prompting LLMs

In the text generation scenario, we choose different backbones for our prompting pipeline BoFiTGen, as they stand out as the most cutting-edge Large Language Models. Our prompt is composed of four sections. For each video V_i, we set up a context description c to give thorough explanations of the following matrix R_i. Next, we append the prompt with a universal question q about the task to be accomplished in its answer. Afterwards, notes n are given to BoFiTGen, specifically on the equipment type, text length, granularity limitation, style of writing, and its persona (i.e. a fitness training coach). Finally, we add the angular representation matrix R_i to the prompt. Overall, the assembled prompt P_i for the zero-shot prompting scenario can be summarized as the string-concatenation of c, q, n, R_i. We denote (R_0, I_0) as a data pair, presented here as an in-context example, where R_0 is the angular representation of the given video and I_0 is its annotated text description. In the one-shot prompting scenario, we can formalize the prompt as $P_i = [c, q, n, R_0, I_0, R_i]$. We obtain $\hat{I}_i$ from $BoFiTGen(P_i)$, representing the generated text description of the given video V_i by BoFiTGen with prompt P_i.

5 Experiment

We evaluate BoFiTGen on its capability of describing fine-grained human motions under zero-shot and one-shot prompting scenarios. Experiments are conducted on BoFiT. Comprehensive evaluation metrics and in-depth implementation details are provided below.

5.1 Metrics

Performance on BoFiT is evaluated according to different metrics that demonstrate the text generation capability. The evaluation metrics used in our experiments are all supervised metrics that compute the text-to-text similarity between generated sentences and reference sentences: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), BERTScore (Zhang et al., 2019) and FCE-Motion. While BERTScore evaluates similarity using word embeddings, the other metrics rely on n-gram-based token matching.

Specifically, FCE is an order-sensitive metric on the evaluation of fine-grained motion description (Yu et al., 2018). In this paper, we only evaluate the accuracy of the verb, formalizing FCE as FCE-Motion. It focuses on human motions and their temporal order in text.

5.2 Implementation details

We compare the fine-grained human motion video captioning ability of different VLMs and BoFiTGen. Note that conventional video captioning methods without the assistance of LLMs have failed to perform well on BoFiT, as the length of generated text is too short for them to gain a reasonable value on evaluation metrics.

In detail, we evaluate the performance of recent VLMs, including Video-LLaMA (Zhang et al., 2023), Video-ChatGPT (Maaz et al., 2023), and Video-LLaVA (Lin et al., 2023). For BoFiTGen, we employ different LLM backbones: LLaMA2-13b (Touvron et al., 2023), LLaMA3.1-8b(Dubey et al., 2024), Vicuna-13b(Chiang et al., 2023), ChatGPT(gpt-3.5-turbo-1106) and Mistral-7b (Jiang et al., 2023). Considering the high cost of inference with GPT4 (Achiam et al., 2023), we only conduct experiments on a subset of BoFiT with 378 samples. Concrete experimental data can be found in Appendix.E. We design different prompts for VLMs and BoFiTGen respectively. For VLMs, we let the model describe the human motion in the video as a professional training coach, limiting the output text length to approximately 130 words, matching the average length of ground truth descriptions. For BoFiTGen, we sample an average of 5 frames from each video and extract intermediate representations from this frame sequence. We then prompt the model to describe the human motion based on the given angle matrix and additional information such as the type of equipment. We measure the results separately for both scenarios that utilize different angle representations in motion data modeling. We also condition BoFiTGen with the same text length limitation. To minimize distractions caused by the given angle matrix, we instruct the LLMs not to include specific numbers in their response. For all models, we use off-the-shelf pre-trained weights for fast inference, setting the temperature to zero and other parameters to default.

5.3 Performance Analysis

We evaluate the prompting performances of each model under the combination of two factors. One factor is the deployment of zero-shot or one-shot

Method	Backbone	B@1	B@2	B@3	B@4	R	M	C	FCE-M	BERT
video and prompt inputs										
Video-LLaMA	-	0.167	0.052	0.017	0.006	0.156	0.079	0.003	0.198	-0.086
Video-LLaVA	-	0.356	0.181	0.099	0.061	0.24	0.177	0.023	0.407	0.173
Video-ChatGPT	-	0.212	0.089	0.043	0.023	0.172	0.092	0.010	0.311	0.081
prompt inputs only (zero-shot)										
BoFiTGen-inc	LLaMA2-13B	0.274	0.143	0.077	0.047	0.225	0.173	0.016	0.338	0.085
	LLaMA3.1-8B	0.278	0.137	0.070	0.039	0.230	0.143	0.020	0.364	0.162
	Vicuna-13B	0.319	0.172	0.099	0.061	0.242	0.156	0.031	0.409	0.215
	ChatGPT	0.295	0.150	0.077	0.044	0.234	0.174	0.016	0.372	0.153
	Mistral-7B	0.262	0.138	0.076	0.046	0.213	0.195	0.003	0.323	0.082
BoFiTGen-tb	LLaMA2-13B	0.129	0.053	0.023	0.012	0.145	0.060	0.006	0.143	-0.072
	LLaMA3.1-8B	0.295	0.147	0.079	0.048	0.231	0.145	0.023	0.359	0.150
	Vicuna-13B	0.279	0.145	0.081	0.050	0.226	0.135	0.013	0.395	0.181
	ChatGPT	0.293	0.163	0.097	0.063	0.245	0.148	0.018	0.422	0.236
	Mistral-7B	0.254	0.127	0.068	0.041	0.216	0.186	0.002	0.314	0.104
prompt inputs only (one-shot)										
BoFiTGen-inc	LLaMA2-13B	0.327	0.185	0.109	0.070	0.256	0.164	0.031	**0.448**	0.253
	LLaMA3.1-8B	0.299	0.147	0.079	0.048	0.231	0.146	0.023	0.359	0.150
	Vicuna-13B	0.354	0.201	0.120	0.076	0.259	0.172	0.036	0.429	0.213
	ChatGPT	**0.384**	**0.219**	**0.130**	**0.083**	**0.272**	0.183	**0.068**	0.438	0.257
	Mistral-7B	0.286	0.156	0.089	0.055	0.228	**0.199**	0.005	0.355	0.090
BoFiTGen-tb	LLaMA2-13B	0.351	0.184	0.104	0.065	0.228	0.177	0.021	0.380	0.193
	LLaMA3.1-8B	0.354	0.194	0.111	0.069	0.252	0.171	0.053	0.396	0.214
	Vicuna-13B	0.328	0.177	0.105	0.067	0.248	0.159	0.034	0.42	0.191
	ChatGPT	0.340	0.192	0.116	0.074	0.253	0.169	0.040	0.438	**0.270**
	Mistral-7B	0.374	0.206	0.120	0.074	0.253	0.190	0.065	0.415	0.240

Table 2: The BLEU (B), ROUGE-L (R), METEOR (M), CIDEr (C), FCE-Motion (FCE-M) and BERTScore (BERT) of LLMs, where inc refers to included angle representation and tb refers to Tait-Bryan angle representation.

prompting methods, and the other is the utilization of included angle representation or Tait-Bryan angle representation. From Table 2, we find that:

In general, VLMs perform worse than BoFiTGen. Due to the fact that VLMs lack the ability to align images and pose semantics at a fine-grained level, they make mistakes in generating detailed descriptions. Prompted with angle matrices, BoFiTGen can directly analyze the pose information, but still have difficulties understanding complex intermediate representations. This may be the reason why some LLMs perform worse than VLMs in zero-shot prompting scenarios. However, the performance of BoFiTGen is promoted greatly in one-shot prompting scenarios, showing superiority over VLMs. A case study comparing the SOTA VLM (i.e. Video-LLaVA) with BoFiTGen is shown in Appendix.F.

One-shot prompting results in better performance than zero-shot prompting. With the guidance of in-context examples, LLMs can better learn the correspondence between the intermediate representations and human motion descriptions. Among them, LLaMA2-13B using the Tait-Bryan angle representation improves the most. ChatGPT performs best in most of the settings.

There is no singular best angle representation system across all backbone models. Overall, Mistral-7B and LLaMA3.1-8B performs better with the Tait-Bryan angle representation, while other models perform better with the included angle representation. LLaMA2-13B is particularly bad at understanding Tait-Bryan angles with zero-shot prompting.

5.4 Frame sampling

We evaluate the changes in FCE-Motion and METEOR scores on BoFiTGen with a ChatGPT backbone, with number of frames sampled ranging from 5 to 10. The experiment is conducted on a smaller subset of BoFiT with 378 samples. Results are shown in Figure 5. We find that the FCE-Motion score increases slightly with the increase of frame numbers when using included angle representations, indicating a better motion description capability. Meanwhile, the method using the Tait-Bryan angle representation does not show the same trend. This may result from the Tait-Bryan angle

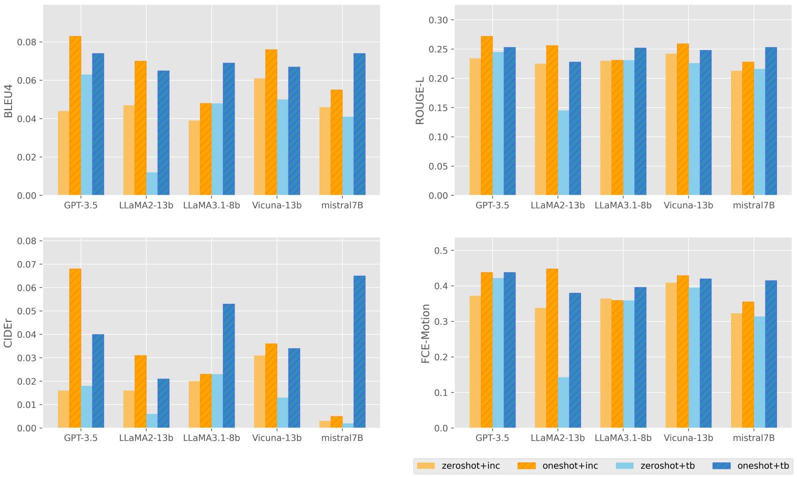


Figure 4: Results under different combinations of angle representations and prompting methods. For example, "zeroshot+inc" means this BoFiTGen baseline employs included angle representation and zero-shot prompting.

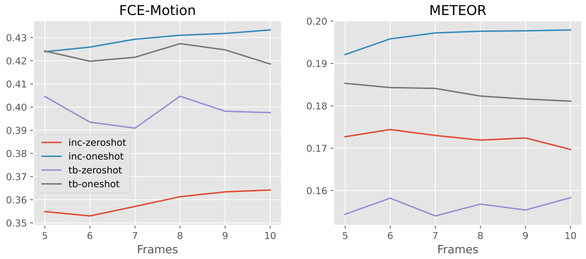


Figure 5: Visualizations of the relationship between evaluation metrics and frame numbers.

representation's longer prompt, which increases greatly with the rising number of frames and distract LLM's attention on the variance of sequential data.

5.5 Time-disordered investigation

We randomly shuffle the input angle sequences throughout the timeline. Under this operation, all methods' FCE-Motion scores decrease with an average of 1.7%. Figure 6 shows a case example generated for the same video before and after shuffling. The two descriptions deliver different motions respectively, indicating that our method is sensitive to time order. However, the presence of overlapping verbs from two descriptions results in a small difference in the chronological accuracy of the verbs, thus leading to the relatively subtle decrease in the FCE-Motion score. This suggests more room for improvement in the order-sensitive evaluation metrics for motion description.

Ordered
Stand upright with your feet shoulder-width apart. Ensure that your head is straight and your shoulders are back. Bend your knees slightly while keeping your back straight. Lower your body into a squat position, ensuring that your knees do not extend past your toes. Return to the starting position by pushing through your heels and straightening your legs. Repeat the squat motion for the recommended number of repetitions.

Shuffled
Stand with your feet shoulder-width apart and your arms relaxed at your sides. Bend your knees slightly and tilt your pelvis forward to engage your core muscles. Keep your spine straight and your head aligned with your spine, looking forward. Shift your weight onto one leg and simultaneously lift the opposite leg off the ground, extending it straight back. Maintain a slight bend in your supporting knee and engage your glutes to keep your leg lifted. Hold this position for a few seconds and then switch to the other leg. Repeat for the recommended number of repetitions, alternating legs each time.

Figure 6: Descriptions generated for the same video before and after shuffling.

5.6 Finetuning LLaMA2-13B

Although off-the-shelf pre-trained LLMs are accessible and practicable in comprehensive applications, their performance still suffer from limited generalization ability to unseen tasks. Therefore, to examine the usability of our dataset, we further finetune an open-source model on BoFiT and compare its performance with the original one under the same experimental settings. We use LLaMA2-13B-chat as the baseline model and finetune it using LoRA (Hu et al., 2021) on $(instruction, output)$ pairs. We obtain all instructions following Tait-Bryan angle representation and zero-shot prompting method, and use annotations in BoFiT as output. Detailed settings are shown in Appendix.D. As Table 3 shows, all metrics raise greatly, which demonstrates the potential of BoFiT on finetuning models for such tasks.

6 Conclusions

We construct BoFiT, a fine-grained fitness training dataset for video captioning. We also propose

Model	B@4	R	M	C	FCE-M	BERT
LLaMA2-13B	0.016	0.151	0.080	0.006	0.224	0.074
LLaMA2-13B^f	0.087↑	0.265↑	0.181↑	0.137↑	0.460↑	0.265↑

Table 3: The performance of LLaMA2-13B before and after finetuning. LLaMA2-13B^f refers to the finetuned model.

BoFiTGen, a generic method that converts human motion to textual prompts and generates video captions via LLM. Through experiments under zero-shot and one-shot scenarios, we find that BoFiTGen outperforms previous VLMs on BoFiT on comprehensive metrics. In our opinion, BoFiTGen reveals that LLMs implicitly have the ability to understand pose encoding, which provides a new possibility for both video captioning and LLMs. Furthermore, we aim to improve the data quality of BoFiT in the future, hence contributing to the community with a more reliable evaluation benchmark on the fine-grained video captioning task.

Limitations

To our best knowledge, we are the first to propose the fine-grained human motion video captioning task. Since it is difficult to manually develop a corresponding dataset, we acquire annotated pairs of videos and their descriptions from the Internet. Specifically, we supplement missing annotations with the assistance of LLM. The quality assurance work is done through human evaluation. However, due to the limitation of human resources, we only evaluate on a sampled subset of the full dataset. Therefore, we would like to contribute more efforts in promoting the quality of the BoFiT dataset in the future. In addition, since we make use of human pose features as intermediate representations between video and text, it may lead to some information loss, such as specific movements of the equipment. We will continue to explore more reasonable intermediate representations to help LLMs understand videos better, as this challenge may also be open to other researchers in this field.

Acknowledgements

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Key R&D Program of China (No. 2022YFC3600402).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ijaz Akhter and Michael J Black. 2015. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Paul Berner, Ralph Toms, Kevin Trott, Farid Mamaghani, David Shen, Craig Rollins, and Edward Powell. 2008. Technical concepts orientation, rotation, velocity and acceleration, and the srm. *TENA (Test & Training Enabling Architecture) project by SEDRIS*, 21.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2024. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. 2024. Prollm: Protein chain-of-thoughts enhanced llm for protein-protein interaction prediction. *arXiv preprint arXiv:2405.06649*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Lavender: Unifying video-language understanding as masked language modeling. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23119–23129.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *ArXiv*, abs/2311.10122.

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

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Swinbert: End-to-end transformers with sparse attention for video captioning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17928–17937.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv*, abs/2306.05424.

Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. 2019. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10132–10141.

Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*.

Bo Pang, Kaiwen Zha, Hanwen Cao, Chen Shi, and Cewu Lu. 2018. Deep rnn framework for visual sequential applications. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 423–432.

Bo Pang, Kaiwen Zha, Hanwen Cao, Jiajun Tang, Minghui Yu, and Cewu Lu. 2020. Complex sequential understanding through the awareness of spatial and temporal concepts. *Nature Machine Intelligence*, 2:245 – 253.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. *arXiv preprint arXiv:1809.04560*.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762.

Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. 2019. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17938–17947.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. 2020. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Alessandro Suglia, José Lopes, Emanuele Bastianelli, Andrea Vanzo, Shubham Agarwal, Malvina Nikandrou, Lu Yu, Ioannis Konstas, and Verena Rieser. 2022. Going for goal: A resource for grounded football commentaries. *arXiv preprint arXiv:2211.04534*.

Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. 2017. Integral human pose regression. *ArXiv*, abs/1711.08229.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. 2021. Clip4caption: Clip for video caption. *Proceedings of the 29th ACM International Conference on Multimedia*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2022. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2022. Hitea: Hierarchical temporal-aware video-language pre-training. *ArXiv*, abs/2212.14546.

Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. 2018. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6006–6015.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435.

Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. 2023. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665.

Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. 2019. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2344–2353.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2017. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023b. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Appendix

A The Definition Details of the Included Angle Representation

Type	Associated Body Trunks	Definition	Explanation
Local	upper body	angle between Z-axis and the ground	Standing/Lying
	upper body	angle between the upper body and Z-axis	Pitching
	left(right) upper arm	angle between the left(right) upper arm and X-axis	Moving left and right
	left(right) upper arm	angle between the left(right) upper arm and Y-axis	Moving front and back
	left(right) upper arm	angle between the left(right) upper arm and Z-axis	Moving up and down
	left(right) elbow	angle between the left(right) upper arm and the forearm	Flexing/Extending
	left(right) thigh	angle between the left(right) thigh and X-axis	Moving left and right
	left(right) thigh	angle between the left(right) thigh and Y-axis	Moving front and back
	left(right) thigh	angle between the left(right) thigh and Z-axis	Moving up and down
	left(right) knee	angle between the left(right) thigh and the lower leg	Flexing/Extending
Global	full body	twisted or rotated angle of the upper body	Rotating
	full body	distance of feet off the ground	Jumping
	full body	distance moved forward relative to the initial state	Translating
	full body	distance moved leftward relative to the initial state	Translating

Table 4: The definition of the included angle representation. Note that the X, Y, and Z axes are according to the definition of the human coordinate system.

A.1 Local Motion Representation

For a joint with only one DOF (degree of freedom), we calculate the angle between the two rigid bodies connected to this joint. Otherwise, we calculate the angles that the body trunk forms with the three axes of the human coordinate system. Here we offer two examples of joints with different degrees of freedom: For the knee with DOF=1, we use the angle between the thigh and the lower leg to represent the flexion/extension. For the hip with DOF=3, we use the angle of the thigh to the three axes mentioned above to represent its movement. For necessary simplification, we ignore some subtle rotations made by the wrists and ankles.

A.2 Global Motion Representation

Global clues provided to LLMs are represented as the rotation angles of the Y-axis, the distance of feet off the ground, the distance of the forward translation, and the distance of leftward translation. For each video frame, the above data is calculated from the distance to the initial frame.

B Quaternions to Tait-Bryan Angles

$$\boldsymbol{q} = q_0 + q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} \tag{1}$$

$$\phi = \operatorname{arctan2}\left(q_2q_3 + q_0q_1, \frac{1}{2} - ({q_1}^2 + {q_2}^2)\right) \tag{2}$$

$$\theta = \arcsin\left(-2(q_1q_3 - q_0q_2)\right) \tag{3}$$

$$\psi = \operatorname{arctan2}\left(q_1q_2 + q_0q_3, \frac{1}{2} - ({q_2}^2 + {q_3}^2)\right) \tag{4}$$

Given quaternions $\boldsymbol{q}$, the Tait-Bryan angles ϕ, θ, ψ are computed by Eq.1 to 4.

C Prompts of BoFiTGen

The prompts of BoFiTGen consist of 5 parts, named as 1) context description, 2) instruction, 3) notes, 4) in-context example, 5) equipment and motion matrix. The in-context example is only introduced in one-shot scenarios. Detailed prompts for each part are presented as follows:

C.1 Context Description and Instruction

Description and Instruction of Included Angle Representation

```
I will provide you with the information of a bodybuilding exercise, including the
equipment used in it and a matrix of (5, 22), The first dimension of the matrix
represents the 5 time points in the process of the action, including the initial
posture. Each time point has a vector of size 22 to describe the human posture at
this node. Respectively, the 22 numbers in the vector of a time point are: [Angle
between top of the body and ground, angle between left upper arm and top of the
body, angle between left upper arm and front of the body, angle bewteen left upper
arm and right of the body, angle of left elbow, angle between right upper arm and
top of the body, angle between right upper arm and front of the body, angle
between right upper arm and right of the body, angle of right elbow, angle between
upper body and top of the body, angle between left thigh and top of the body,
angle between left thigh and front of the body, angle between left thigh and right
of the body, angle of left knee, angle between right thigh and top of the body,
angle between right thigh and front of the body, angle between right thigh and
right of the body, angle of right knee, twist or rotate angle of upper body,
distance off the ground with feet, distance moved forward relative to initial
state, distance moved to left relative to initial state].
The direction from left hip to right hip is defined as the right of the body, the
direction from middle point of the pelvis to lumbar spine is defined as the top of
the body. Perpendicular to them is the front of the body.
Please give an instruction of the action step by step based on the matrix
information.
```

Description and Instruction of Tait-Bryan Angle Representation

```
I will provide you with the information of a bodybuilding exercise, including the
equipment used in it and a matrix of size (5, 17, 3). The first dimension of the
matrix represents the 5 time points in the process of the action, including the
initial posture. Each time point has a matrix of size (17, 3) to describe the
human posture at this node. The first row is the coordinates of the pelvis at this
moment in the global coordinate system, the second row is the global rotation of
the pelvis (root node) in the global coordinate system, and the third to the
seventeenth rows represent the tait-bryan angles representation of the relative
rotation of the 15 human joints relative to their parent node in the kinematic
tree. The three numbers in each row are the angles of yaw, pitch, and roll in
degrees. The joints are as follows:
Row 3: Left Hip, Row 4: Right Hip, Row 5: Left Knee, Row 6: Right Knee, Row 7:
Spine, Row 8: Left Ankle, Row 9: Right Ankle, Row 10: Neck, Row 11: Head, Row 12:
Left Shoulder, Row 13: Right Shoulder, Row 14: Left Elbow, Row 15: Right Elbow,
Row 16: Left Wrist, Row 17: Right Wrist.
Please give an instruction of the action step by step based on the matrix
information.
```

C.2 Notes

```
Note:
1. You should infer the global posture information of the human body from the
global coordinate system, such as standing, supine, etc.
2. The text you provide should be around 130 words.
3. Do not include specific angles or coordinates in the description, but infer the
movement of the relevant parts of the human body based on the angle information.
4. Please provide only the descriptive text, no extra words.
5. Imagine yourself as a bodybuilding coach, how would you teach others to perform
this action step by step? But do not have too many steps in the description, four
or five steps are enough.
```

C.3 In-context Example

```
The following Question 1 is an example. Please answer Question 2.
Question 1: The equipment used is {EXAMPLE EQUIPMENT}. The matrix is: {EXAMPLE
MOTION MATRIX}
Answer 1: {EXAMPLE ANNOTATION}
```

C.4 Equipment and Motion Matrix

Zero-shot Setting

```
Question: The equipment used is {EQUIPMENT}. The matrix is: {MOTION MATRIX}
Answer:",
```

One-shot Setting

```
Question 2: The equipment used is {EQUIPMENT}. The matrix is: {MOTION MATRIX}
Answer 2:",
```

D Settings of Finetuning LLaMA2-13B

Base Model We use LLaMA2-13B-chat (Touvron et al., 2023) as the base model, which is optimized for instruction understanding using supervised finetuning and RLHF. Since the data scale of BoFiT is not large, finetuning on the chat model can lead to faster convergence than the base model.

Training Data We set the subset of BoFiT with 378 samples as the test set, and the remaining data are used as the train set. In the train set, we randomly select 20 samples as a small dev set.

Settings We use LoRA to finetune the base model, where the LoRA rank is set to 8. Using a max sequence length of 2048, a batch size of 32 and a learning rate of 1e-4, we finetune the model for 3000 steps with a 100-step warmup. We set the micro-batch size to 2, and finish the training on a single NVIDIA A800 GPU in about one day. For inference, we select the best checkpoint based on the minimal loss on the dev set.

E Evaluation on the Sampled Subset

E.1 The Statistics of the Sampled Subset

Equipment Type	Number of Videos
body-only	149
dumbbells	79
barbells	47
kettlebells	34
others	69
overall	378

Table 5: The equipment composition of the subset.

E.2 Experimental Results on the Sampled Subset

Method	Backbone	B@1	B@2	B@3	B@4	R	M	C	FCE-M	BERT
video and prompt inputs										
Video-LLaMA	-	0.172	0.054	0.018	0.007	0.162	0.092	0.005	0.247	0.028
Video-ChatGPT	-	0.198	0.088	0.045	0.026	0.185	0.110	0.019	0.339	0.102
Video-LLaVA	-	0.288	0.136	0.071	0.041	0.211	0.132	0.030	0.357	0.172
prompt inputs only										
BoFiTGen-inc	LLaMA2-13B	0.276	0.142	0.076	0.046	0.224	0.171	0.014	0.345	0.131
	Vicuna-13B	0.347	0.183	0.103	0.063	0.243	0.166	0.055	0.385	0.219
	ChatGPT	0.321	0.172	0.095	0.058	0.248	0.175	0.048	0.365	0.174
	GPT-4	0.308	0.140	0.059	0.027	0.227	0.150	0.052	0.326	0.179
BoFiTGen-tb	LLaMA2-13B	0.173	0.064	0.029	0.016	0.151	0.080	0.006	0.224	0.074
	Vicuna-13B	0.347	0.183	0.103	0.063	0.243	0.166	0.055	0.385	0.187
	ChatGPT	0.326	0.173	0.099	0.063	0.250	0.158	0.031	0.406	0.195
	GPT-4	0.320	0.144	0.065	0.033	0.227	0.161	0.060	0.348	0.174

Table 6: The BLEU (B), ROUGE-L (R), METEOR (M), CIDEr (C), FCE-Motion (FCE-M), and BERTScore (BERT) of VLMs and LLMs in the zero-shot prompting scenario, where inc refers to included angle representation and tb refers to Tait-Bryan angle representation.

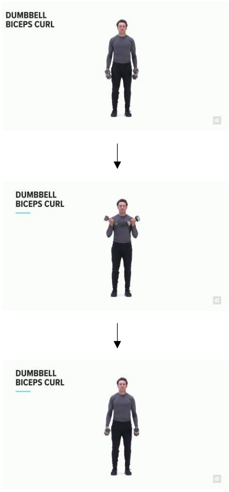
Method	Backbone	B@1	B@2	B@3	B@4	R	M	C	FCE-M	BERT
prompt inputs only										
BoFiTGen-inc	LLaMA2-13B	0.370	0.206	0.120	0.076	0.257	0.176	0.056	0.418	0.239
	Vicuna-13B	0.374	0.212	0.127	0.083	0.264	0.186	0.078	0.407	0.213
	ChatGPT	0.402	0.231	0.139	0.090	0.277	0.192	0.090	0.436	0.258
	GPT-4	0.349	0.171	0.084	0.045	0.241	0.172	0.074	0.373	0.214
BoFiTGen-tb	LLaMA2-13B	0.337	0.184	0.107	0.067	0.244	0.185	0.043	0.392	0.222
	Vicuna-13B	0.361	0.195	0.115	0.074	0.253	0.172	0.055	0.418	0.207
	ChatGPT	0.385	0.220	0.134	0.088	0.262	0.184	0.079	0.432	0.275
	GPT-4	0.334	0.167	0.086	0.048	0.240	0.183	0.050	0.374	0.219

Table 7: The BLEU (B), ROUGE-L (R), METEOR (M), CIDEr (C), FCE-Motion (FCE-M), and BERTScore (BERT) of LLMs in the one-shot prompting scenario, where inc refers to included angle representation and tb refers to Tait-Bryan angle representation.

We also conduct an experiment on a subset of 378 samples. This is to take the performance of GPT4 into consideration. Since GPT4 has emerged as the most up-to-date Large Language Model, we wonder how it performs on the fine-grained human motion captioning task. To our surprise, it only ranked in the middle among all baselines. The results are presented in Table 6 and Table 7.

According to Gruver et al. (2024), GPT-4 performs badly for time-series prediction, due to the fact that it is an expert in alignment instead of numerical sequence processing. This may explain the weaker performance of GPT-4 compared to other less-advanced models.

F Case Study



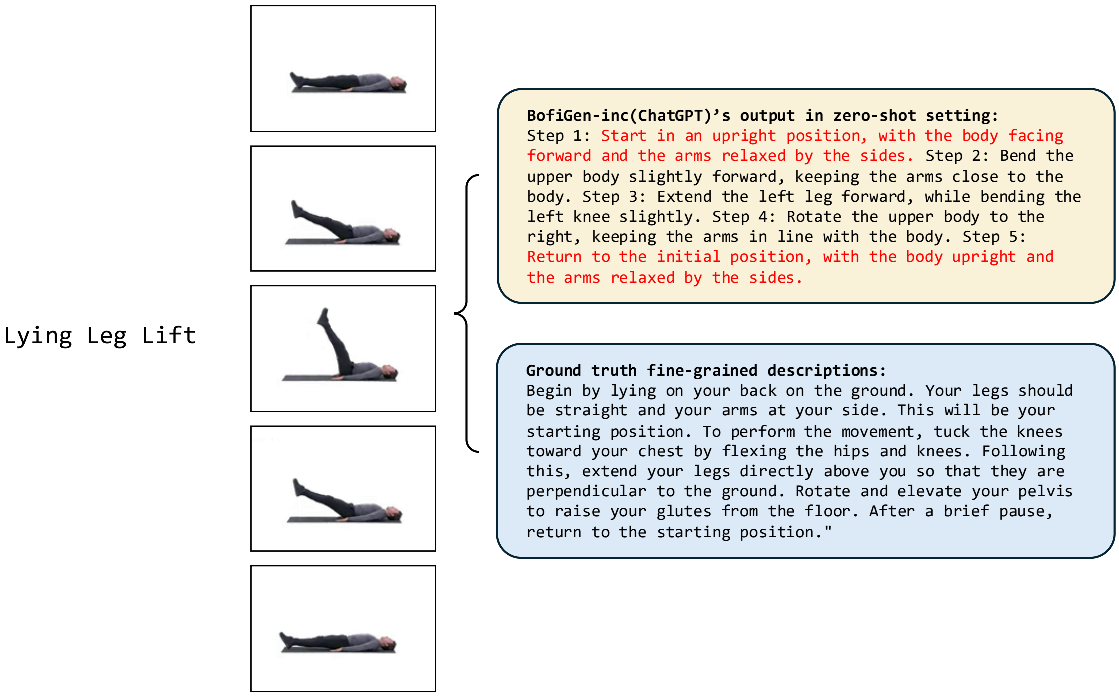
Ground truth	Video-LLaVA	Ours (zero-shot)	Ours (one-shot)
Stand up straight with a dumbbell in each hand at arm's length. Keep your elbows close to your torso and rotate the palms of your hands until they are facing forward. This will be your starting position. Now, keeping the upper arms stationary, exhale and curl the weights while contracting your biceps. Continue to raise the weights until your biceps are fully contracted and the dumbbells are at shoulder level. Hold the contracted position for a brief pause as you squeeze your biceps. Then, inhale and slowly begin to lower the dumbbells back to the starting position. Repeat for the recommended amount of repetitions.	The video is about a man demonstrating how to do a bodybuilding exercise. He starts by standing with his feet shoulder-width apart and holding a set of weights in his hands. He then bends his knees slightly and lifts the weights up to his shoulders, keeping his back straight. He then lifts the weights over his head, extending his arms fully. Finally, he lowers the weights back down to his shoulders and repeats the exercise.	Step 1: Stand upright with feet shoulder-width apart, holding a dumbbell. Step 2: Bend your knees slightly while keeping your back straight and pelvis aligned. Step 3: Pivot at the hips while maintaining a neutral spine, bringing the dumbbell down toward the floor. Step 4: Lower the dumbbell between your legs, feeling a stretch in your hamstrings and glutes. Step 5: Drive through your heels and extend your hips, bringing the dumbbell back to the starting position. Repeat for desired reps.	Start by holding a dumbbell in each hand with your palms facing inwards and your arms extended at your sides. Keep your feet shoulder-width apart. Bend your knees slightly and lean forward at the waist while keeping your back straight. This will be your starting position. While keeping your upper body stationary, lift the dumbbells by raising your elbows and bringing them up to shoulder level. Exhale as you perform this movement. Hold the contraction for a second before slowly lowering the dumbbells back to the starting position while inhaling. Repeat for the recommended amount of repetitions.

Figure 7: Comparison between text generated by VLM and BoFiTGen based on zero-shot and one-shot prompts. Results of BoFiTGen are generated by ChatGPT using Tait-Bryan angle representation.

Figure 7 shows a video sample and its captions generated by different models. Both the Video-LLaVA and our zero-shot methods contain factual errors, while one-shot results are of significantly higher quality. Although the only information we provide about the equipment is its name, LLM could infer the location of the equipment by its intrinsic reasoning capability.

G Failure Cases

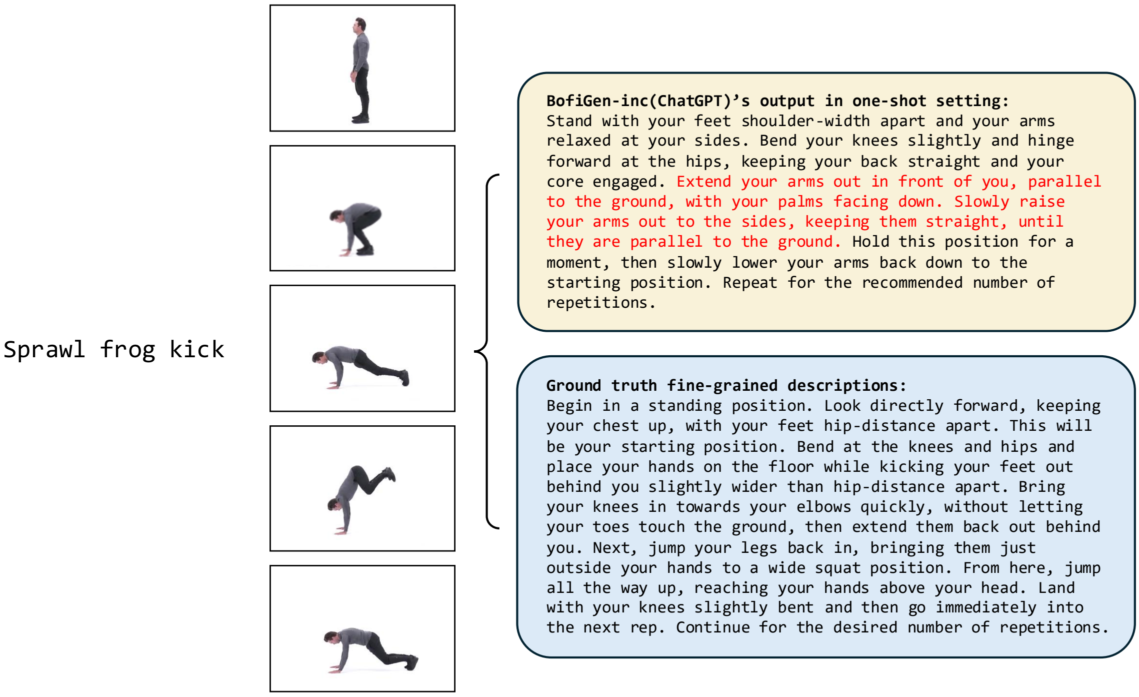
Figure 8 and Figure 9 show the failure cases of BoFiTGen-inc in both zero-shot and one-shot settings. The two cases all involve actions that must be performed with the body parallel to the ground, demonstrating the shortcomings of BoFiTGen in understanding and explaining non-standing positions.



BofiGen-inc(ChatGPT)'s output in zero-shot setting:
Step 1: Start in an upright position, with the body facing forward and the arms relaxed by the sides. Step 2: Bend the upper body slightly forward, keeping the arms close to the body. Step 3: Extend the left leg forward, while bending the left knee slightly. Step 4: Rotate the upper body to the right, keeping the arms in line with the body. Step 5: Return to the initial position, with the body upright and the arms relaxed by the sides.

Ground truth fine-grained descriptions:
Begin by lying on your back on the ground. Your legs should be straight and your arms at your side. This will be your starting position. To perform the movement, tuck the knees toward your chest by flexing the hips and knees. Following this, extend your legs directly above you so that they are perpendicular to the ground. Rotate and elevate your pelvis to raise your glutes from the floor. After a brief pause, return to the starting position."

Figure 8: BoFiTGen-inc's failure case with ChatGPT in zero-shot setting



BofiGen-inc(ChatGPT)'s output in one-shot setting:
Stand with your feet shoulder-width apart and your arms relaxed at your sides. Bend your knees slightly and hinge forward at the hips, keeping your back straight and your core engaged. Extend your arms out in front of you, parallel to the ground, with your palms facing down. Slowly raise your arms out to the sides, keeping them straight, until they are parallel to the ground. Hold this position for a moment, then slowly lower your arms back down to the starting position. Repeat for the recommended number of repetitions.

Ground truth fine-grained descriptions:
Begin in a standing position. Look directly forward, keeping your chest up, with your feet hip-distance apart. This will be your starting position. Bend at the knees and hips and place your hands on the floor while kicking your feet out behind you slightly wider than hip-distance apart. Bring your knees in towards your elbows quickly, without letting your toes touch the ground, then extend them back out behind you. Next, jump your legs back in, bringing them just outside your hands to a wide squat position. From here, jump all the way up, reaching your hands above your head. Land with your knees slightly bent and then go immediately into the next rep. Continue for the desired number of repetitions.

Figure 9: BoFiTGen-inc's failure case with ChatGPT in one-shot setting