Exploring Fine-Grained Human Motion Video Captioning

Anonymous ACL submission

Abstract

Fine-grained human motion descriptions are crucial for people's fitness training as well as their health management. Naturally, it brings the problem of fine-grained human motion video-to-text generation into our focus. Previous video captioning models, including LLM-driven methodologies, are short of capturing fine-grained semantics of the videos through modeling. Meanwhile, the generated descriptions are brief and lack fine details in demonstrating human motion. Hence, existing methods driven by short and coarse-grained ground-truth descriptions still have room for improvement, given the fact that datasets with fine-grained, annotated long text are in deficiency.

In this paper, we construct a fine-grained motion video captioning dataset named BoFiT (**Bo**dy **Fi**tness **T**raining), which is composed of fitness training videos, paired with human motion descriptions temporally at step granularity and spatially at body-trunk granularity. We also implement a state-of-the-art baseline named PoseGPT, with the assistance of the 3D Human Pose Estimation model, MotionBERT. It extracts angular representations of the videos and encodes them into prompts. These prompts are later used by LLMs to generate fine-grained descriptions of human motions.

Results show that PoseGPT outperforms other previous methodologies 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.

1 Introduction

Nowadays, with the increasing pressure of modern life, people turn to find ways to keep fit and stay healthy at the fast pace of living. They tend to work out in gyms or at home while seeking tutorship in fitness channels and apps. However, self-training video courses raise a challenge: trainers may not know exactly how to follow the video in detail and how well they act in repeating them. To make fitness training more accurate, reliable, and inexpensive, we need fine-grained human motion descriptions generated from motion videos.

The existing datasets of human motion videos are widely 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). This kind of ground truth caption of a video is of keyword level, far from the fine-grained (i.e. step-by-step, body trunk level descriptive text for instructional purposes) human motion descriptions. Later on, a series of specific sports video datasets have been constructed, falling in domains ranging from basketball, volleyball, and football competitions (Yu et al., 2018; Pasunuru and Bansal, 2018; Qi et al., 2019; Suglia et al., 2022). To the best of our knowledge, these datasets are developed mainly from the human interaction level but do not focus on the fine-grained motions of body trunks. Hence we propose a novel task called fine-grained human motion video captioning to fill in the blanks of previous works.

Motivated by this, we need to construct a corresponding dataset. However, it is hard to develop a human motion video dataset with fine-grained captions. On the one hand, as we require professional fitness training videos, the expertise of the recorded trainer is highly demanded. On the other hand, the annotation of the ground truth captions consumes a huge workload and could suffer from discrepancies in the granularity of the descriptions due to human subjectivity. To tackle the above difficulties, we build a dataset named BoFiT (Body Fitness Training Dataset), sourced from BodyBuilding[1] since it has legible and professional training videos with fine-grained, body-trunk level descriptions. Specifically, we supplemented some incomplete

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

descriptions of the data using LLM and manual proofreading methods.

As videos in Bodybuilding are paired with fine-grained long texts, previous video-to-text methods that are short in the capability of long text generation do not fit in this scenario (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). Since LLMs are skilled at the above task, LLM-based methods naturally become the mainstream solution to this task. Existing multimodal Large Language Models like Video-ChatGPT (Maaz et al., 2023), Video-LLaMA (Zhang et al., 2023) and Video-LLaVA (Lin et al., 2023) are considered cutting-edge methodologies of video captioning in long-text generation scenarios. However, they still underperform on BoFiT by giving wrong depictions of human motions. In this paper, we propose a few-shot LLM method PoseGPT to accomplish the introduced fine-grained human motion video captioning task. In PoseGPT, we first convert human motion videos into intermediate explainable representations to exploit LLMs' powerful ability to analyze, understand, and depict video content at the human-trunk level granularity. Based on the BoFiT dataset, we conduct in-depth experiments to investigate the performance of PoseGPT and other video captioning models on different aspects. The results show that PoseGPT 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 to use human posture features as intermediate representations between video and text, helping large language models well understand videos.

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

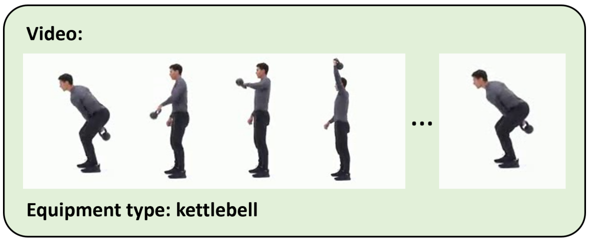
2 Related Work

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 attaches 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 ones to our research focus. On one hand, some existing works formalize dense video captioning as (Krishna et al., 2017) does, aiming at generating short captions for trimmed video clips. Then the overall video would 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 close to our research goal but fail to focus on describing body-trunk-level human motions, generating action-level sports descriptions instead. Here we get deep down into the granularity of human body trunks by constructing BoFiT as a more challenging task than before.

2.2 Large Language Models for Multi-modal Tasks

Recently, many works intend to extend LLMs to understand visual inputs including images and videos. The 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 LLM 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 and BLIP-2 (Li et al., 2023) uses Q-Former to adapt visual features for LLM. Subsequently, InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023), and MiniGPT-4 (Zhu et al., 2023a) explore methods for visual instruction tuning and make VLMs more instruction-aware. Video-LLaMA (Zhang et al., 2023), Video-ChatGPT (Maaz et al., 2023), and Video-LLaVA (Lin et al., 2023) extend inputs from



Fine-grained video caption:

```
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.

images to videos.

2.3 3D Human Pose Estimation

3D Human Pose Estimation involves the retrieval of three-dimensional human poses from monocular RGB videos. To solve this classical problem, methods fall into two distinct categories. One is the single-stage solution, which is to extract 3D pose information from the input images directly (Sun et al., 2017; Moon et al., 2019; Zhou et al., 2019). The other one is the two-stage solution, which extracts the 2D poses first and then lifts them to 3D coordinates through a single neural network. Its performance relies heavily on the 2D extractor and the lifting model. The former one has achieved great performance by the combination of the backbone network and the 2D heatmap representation (Simonyan and Zisserman, 2014; He et al., 2015; Newell et al., 2016; Pang et al., 2018, 2020), while the latter one gets advanced through different neural network architectures (Cai et al., 2019; Cheng et al., 2020; Li et al., 2021).

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 capturing the movement of an individual, one model is expected to generate a fine-grained description 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 PoseGPT generates long captions that depict detailed human motion. To accompany the proposed task, we construct a dataset named BoFiT.

3.2 BoFiT Dataset

We collect original videos from BodyBuiding, a professional fitness training instructional website. These videos have been provided with professional information including motion names, short descriptions, benefits, types, equipment, detailed instructions, and so on. To minimize the bias introduced by the vision model, we select those videos featuring a single person exercising with an unobstructed body. We manually select 378 videos and clip each video to contain only one cycle of motion, as the original video may contain several cycles. Then, each clip obtained contains one and only one complete process of one motion.

To equip each video with one fine-grained caption, we first consider getting detailed instructions from the BodyBuilding website. These instructions are of high quality and include detailed descriptions and tips for every motion step. However, only 202 videos have instructions among all 378 videos. For those 176 videos which are not provided with their textual instructions, it is difficult to manually compile professional instructions without expert knowledge in the field of sports. To promote the efficacy of instruction editing, we make use of the strong generation ability of ChatGPT and prompt it to generate instructions. In the prompt, we only provide the motion name for the corresponding video and its expected instruction length, which is set as the average length of existing instructions. This will cause the generated instructions to be independent of video content.

To ensure the consistency between generated instructions and videos, we manually check and revise the instructions. In the same way, we also generate instructions for the 202 videos that already have instructions. To compare the consistency between the LLM-aided instructions and expert instructions, we calculate their ROUGE-L value, which is 0.3526, to some extent verifies the feasibility of our LLM-aided instruction generation method.

Dataset	Scenario	Sentences per second	Words per second
ActivityNet (Heilbron et al., 2015)	Open Domain	0.327	4.410
MSR-VTT (Xu et al., 2016)	Open Domain	0.067	0.621
YouCook2 (Zhou et al., 2017)	Cooking	0.051	0.449
FSN (Yu et al., 2018)	Basketball	0.556	4.901
SVCDV (Qi et al., 2019)	Volleyball	0.366	3.886
PoseGPT	Fitness Training	1.989	33.489

Table 1: Comparisons among video captioning datasets.

Equipment Type	Video Clip Quantity
body-only	149
dumbbells	79
barbells	47
kettlebells	34
others	69
Overall	378

Table 2: Different equipment types and their corresponding video clip quantities in PoseGPT.

3.3 Dataset Statistics

BoFiT has 378 video clips, 2,765 sentences, and 46,458 words in total, where each video clip spans 3.67 seconds on average, paired with 7.3 sentences and 122.9 words on average. The comparison of BoFiT with other video captioning datasets 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 the open domain and sports domain.

In addition to video clips, motion names, and fine-grained descriptions, BoFiT also provides the corresponding equipment information. Different equipment types and their corresponding data quantities are demonstrated in Table 2. In BoFiT, besides sports video clips classified into body-only training, training with dumbbells, barbells, and kettlebells, the remaining videos include other types of equipment, such as bands, plates, medicine balls, etc.

4 Method

We develop a pipeline named PoseGPT. As demonstrated in Figure 2, it first extracts the angular data of the human motion in the given video through a SOTA 3D human pose estimation model, then encodes the data into a carefully designed prompt to generate fine-grained text description through LLM.

4.1 3D Human Pose Estimation

Here we utilize MotionBERT(Zhu et al., 2023b) as the State-Of-The-Art methodology for extracting 3D human motion information from the given videos. On one hand, it can regress the 3D coordinates of human skeleton key points 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 a rudimentary angular representation system named Included Angle Representation that directly computes the angles between different pairs of body limbs, with an assumption of regarding the human body as a composition of rigid bodies.

Firstly, we define a human body coordinate system. The direction from the right hip to the left hip is notated as the Y-axis, the direction from the midpoint of the pelvis to the lumbar vertebrae is notated as the Z-axis, and the direction perpendicular to them is notated as the X-axis.

Then we classify joints into two types according to degrees of freedom. If a joint has only 1 degree of freedom, we only calculate the angle between two rigid bodies connected to the joint. In other cases, we calculate angles between the non-torso rigid body and axes of the human body coordinate system. For example, we use the angle between thighs and calves to represent knees, and angles between thighs and the three axes to represent hips. Notice that we ignore most of the rotations in the included angle representation such as wrists and ankles.

We regard global human motion information as a set of actions: jumping, rotating, and translating. Global clues provided to LLMs separately stand for: the distance of feet off the ground, the rotation angle of the two hips, the distance of the forward

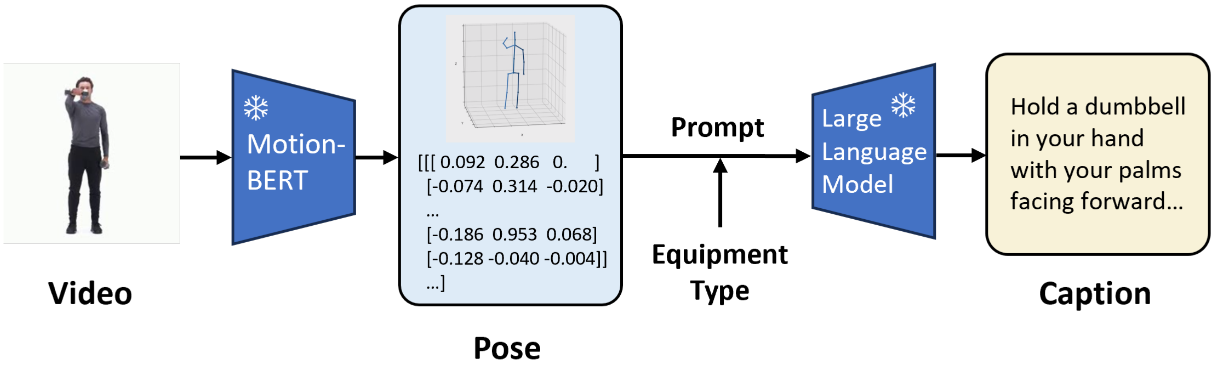


Figure 2: An overview of PoseGPT

translation, and the distance of leftward translation. For each video frame, the above data is calculated from the distance to the initial state.

4.3 Tait-Bryan Angle Representation

We also conduct a more standardized angular modeling system called Tait-Bryan Angle Representation. Normally we define a rotation in the 3D coordinate system as a sequence of three elementary rotations. Specifically, the overall rotation can be factored into the Euler angle convention of three sequential principal rotations. In particular, Tait-Bryan Angles, also known as ZYX Euler Angles, are three sequential rotations made around rotation axis $\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{x}$.

Then we obtain some quaternions predicted by MotionBERT (Zhu et al., 2023b). Since the quaternions depict how each body joint rotates around its precedent on the kinematic tree, we tend to transfer it into a more explainable format. Given the fact that they are trained upon real-world human knowledge (Bubeck et al., 2023), we suppose that Tait-Bryan Angles may serve as a more appropriate resource for prompting LLMs.

According to Berner et al. (2008), if we have a quaternion $\boldsymbol{q} = [q_1, q_2, q_3, q_4]^T$, the Tait-Bryan angles ϕ, θ, ψ are computed by Eq.1 to 3:

$$\phi = \arctan2(q_2 q_3 + q_0 q_1, \frac{1}{2} - ({q_1}^2 + {q_2}^2)) \quad (1)$$

$$\theta = \arcsin(-2(q_1 q_3 - q_0 q_2)) \quad (2)$$

$$\psi = \arctan2(q_1 q_2 + q_0 q_3, \frac{1}{2} - ({q_2}^2 + {q_3}^2)) \quad (3)$$

We generalize the above transformation as the following equation:

$$L_{i,t} = f(Q_{i,t}) \quad (4)$$

In the above equation, i denotes the i^{th} video of BoFiT and t denotes the t^{th} frame. Here $Q_{i,t} \in \mathbb{R}^{16\times 4}$ denotes the local rotation quaternions of the selected 16 human joints (for the pelvis, the root node, is the rotation quaternion in the spatial coordinate system). $f(\cdot)$ denotes the aggregation of the above transformation equations. $L_{i,t} \in \mathbb{R}^{16\times 3}$ denotes the Tait-Bryan angle representations of the same set of rotations. In each row, the three values are the angles of yaw, pitch, and roll in degrees.

The data in BoFiT are first processed by MotionBERT(Zhu et al., 2023b), the current SOTA model in 3D human pose estimation. Note that (V_i, I_i) is a video-text pair. We sample N frames of a given video uniformly on the dimension of time. Let $V_{i,t}$ be the t^{th} frame of the video V_i, for each frame we obtain 3D coordinates and rotation data of the body joints for later use.

$$Q_i = MotionBERT(V_i) \quad (5)$$

At frame t, the local rotation representation matrix $L_{i,t} \in \mathbb{R}^{16\times 3}$ has 16 vectors. Here we add vector $\boldsymbol{g_{i,t}}$ as the global information. It represents the 3D coordinates of the pelvis (i.e. root node) in the global coordinate system. As in Eq.6, we obtain the overall Tait-Bryan representation matrix $R_{i,t}$ by concatenating $\boldsymbol{g_{i,t}}$ and $L_{i,t}$ at the feature dimension.

$$R_{i,t} = [\boldsymbol{g_{i,t}}, L_{i,t}] \quad (6)$$

$$\boldsymbol{g_{i,t}} = [x_r, y_r, z_r] \quad (7)$$

$$L_{i,t,k} = [\alpha, \theta, \phi], k = 1 \ldots 15 \quad (8)$$

As notated in Eq.8, α, θ, ϕ each stands for yaw, pitch, and roll angles as a Tait-Bryan Angle convention of a single rotation. By concatenating $R_{i,t}$ on the dimension of time, for each video V_i, we obtain an overall Tait-Bryan angular representation matrix $R_i \in \mathbb{R}^{N\times 17\times 3}$. The matrix is later used for prompting LLMs for fine-grained human motion description generation.

4.4 Fine-grained Text Generation via Prompting LLMs

In the text generation scenario, we choose different backbones for our prompting pipeline PoseGPT, since they stand out as the most cutting-edge Large Language Models. Comprehensive results are demonstrated in the experiment section.

Our prompt is composed of four sections. Firstly, for each video V_i, we set up a context description c. To give thorough explanations of the provided angular representation matrix R_i, c includes the meaning of each dimension and how they are related to each key point of the human body. Next, we append the prompt with a universal question q about what task to accomplish in its answer. Then, notes n are given to PoseGPT, specifically on the equipment type, text length, granularity limitation, style of writing, and its persona (i.e. a fitness training coach). As Table 2 demonstrates, we provide the equipment types of the fitness training videos since they cannot be distinguished with angular data only. Finally, we add the angular representation matrix R_i to the prompt. Overall, the total prompt P_i for the zero-shot prompting scenario can be summarized as the string-concatenation of c, q, n, R_i, notated as:

$$P_i = [c, q, n, R_i] \tag{9}$$

For the one-shot prompting scenario, we can formalize the prompt as follows:

$$P_i = [c, q, n, R_0, I_0, R_i] \tag{10}$$

In Eq.10, R_0 and I_0 are paired data introduced as an in-context example, where R_0 is the angular representation of the given video and I_0 is its fine-grained text description.

$$\hat{I}_i = PoseGPT(P_i) \tag{11}$$

Here $\hat{I}_i$ denotes the generated fine-grained text description of the given video V_i by PoseGPT with prompt P_i.

5 Experiment

We evaluate our model PoseGPT on its capability of describing fine-grained human motions on zero-shot and one-shot prompting scenarios. The experiments are conducted on PoseGPT, comprehensive evaluation metrics and in-depth implementation details are provided below:

5.1 Metrics

Performance on PoseGPT is evaluated according to different metrics that demonstrate the capability of PoseGPT on the video-to-text task. The evaluation metrics used in our experiments are all supervised metrics that compute the text-to-text similarity between the generated sentences and reference sentences: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and FCE(Yu et al., 2018), an order-sensitive metric on the evaluation of fine-grained motion description. In this paper, we only evaluate the accuracy of the verb in FCE as FCE-Motion, which focuses on human motions and their temporal order.

5.2 Implementation details

In the zero-shot prompting scenario, we comprehensively compare the human motion video captioning ability of different VLMs and PoseGPT, which are implemented with different LLMs. In detail, we evaluate the performance of the recent VLMs, including Video-LLaMA, Video-ChatGPT, and Video-LLaVA. PoseGPT with different LLM backbones (i.e. ChatGPT, GPT-4, Versions of 7B and 13B of LLaMA2 and Vicuna) are all covered in experiments. We separately measure the results of both scenarios that utilize the included angle representation and the Tait-Bryan angle representation in modeling.

We design different prompts for VLMs and PoseGPT through prompt engineering work. For VLMs, we only prompt the model to describe the human motion in the video as a professional bodybuilding coach, with a limited output text length of around 130 words, which is the average length of ground truth descriptions. For PoseGPT, we sample 5 frames from each video uniformly on the timeline and extract angular representations from the frame sequence. Then we prompt the model to describe the human motion according to the given sequence and the provided equipment information. We condition PoseGPT with the same text length limitation. Additionally, to eliminate the negative influence brought by the given angle representations, we let the model not include specific numbers in response. For all models, we utilize off-the-shelf pre-trained weights for fast inference, setting the temperature to zero and other parameters to the default setup.

Method	Backbone	B@1	B@2	B@3	B@4	R	M	C	FCE-M
video and prompt inputs									
Video-LLaMA	-	0.172	0.054	0.018	0.007	0.162	0.092	0.005	0.247
Video-ChatGPT	-	0.198	0.088	0.045	0.026	0.185	0.110	0.019	0.339
Video-LLaVA	-	0.288	0.136	0.071	0.041	0.211	0.132	0.030	0.357
prompt inputs only									
PoseGPT-inc	Llama2-7B	0.281	0.143	0.078	0.048	0.222	0.167	0.024	0.365
	Llama2-13B	0.276	0.142	0.076	0.046	0.224	0.171	0.014	0.345
	Vicuna-7B	0.261	0.143	0.081	0.051	0.235	0.142	0.036	0.359
	Vicuna-13B	0.347	0.183	0.103	0.063	0.243	0.166	0.055	0.385
	ChatGPT	0.321	0.172	0.095	0.058	0.248	**0.175**	0.048	0.365
	GPT-4	0.308	0.140	0.059	0.027	0.227	0.150	0.052	0.326
PoseGPT-tb	Llama2-7B	0.260	0.126	0.066	0.039	0.212	0.154	0.016	0.323
	Llama2-13B	0.173	0.064	0.029	0.016	0.151	0.080	0.006	0.224
	Vicuna-7B	**0.360**	**0.198**	**0.118**	**0.076**	**0.252**	0.172	**0.066**	0.396
	Vicuna-13B	0.347	0.183	0.103	0.063	0.243	0.166	0.055	0.385
	ChatGPT	0.326	0.173	0.099	0.063	0.250	0.158	0.031	**0.406**
	GPT-4	0.320	0.144	0.065	0.033	0.227	0.161	0.060	0.348

Table 3: The BLEU (B), ROUGE-L (R), METEOR (M), CIDEr (C), and FCE-Motion (FCE-M) scores 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
prompt inputs only									
PoseGPT-inc	Llama2-7B	0.298	0.156	0.088	0.054	0.225	0.182	0.024	0.365
	Llama2-13B	0.370	0.206	0.120	0.076	0.257	0.176	0.056	0.418
	Vicuna-7B	0.366	0.203	0.120	0.077	0.253	0.174	0.044	0.408
	Vicuna-13B	0.374	0.212	0.127	0.083	0.264	0.186	0.078	0.407
	ChatGPT	**0.402**	**0.231**	**0.139**	**0.090**	**0.277**	**0.192**	**0.090**	**0.436**
	GPT-4	0.349	0.171	0.084	0.045	0.241	0.172	0.074	0.373
PoseGPT-tb	Llama2-7B	0.305	0.164	0.094	0.059	0.232	0.190	0.022	0.371
	Llama2-13B	0.337	0.184	0.107	0.067	0.244	0.185	0.043	0.392
	Vicuna-7B	0.308	0.170	0.101	0.065	0.251	0.189	0.052	0.410
	Vicuna-13B	0.361	0.195	0.115	0.074	0.253	0.172	0.055	0.418
	ChatGPT	0.385	0.220	0.134	0.088	0.262	0.184	0.079	0.432
	GPT-4	0.334	0.167	0.086	0.048	0.240	0.183	0.050	0.374

Table 4: The BLEU (B), ROUGE-L (R), METEOR (M), CIDEr (C), and FCE-Motion (FCE-M) scores of LLMs in the one-shot prompting scenario, where inc refers to included angle representation and tb refers to Tait-Bryan angle representation.

5.3 Zero-shot performance

We first evaluate the zero-shot prompting performance of each model on PoseGPT, results are shown in Table 3. On most evaluation metrics, PoseGPT pipelines based on different LLM backbones, including LLaMA2, Vicuna, ChatGPT, and GPT-4, outperform Video-LLaVA, the current SOTA methodology on video captioning. Though much less information is provided in prompts compared with video inputs, PoseGPT demonstrates better capabilities in describing fine-grained human motion in videos. Specifically, we find that LLaMA2 struggles with understanding motion when given the Tait-Bryan angle representations. Among them, LLaMA2-13B frequently requests additional information and fails to generate motion descriptions directly with the given prompt inputs. Overall, the pipeline implemented with Vicuna-7B and Tait-Bryan angle representation performs best on most evaluation metrics. Counter-intuitively, the most cutting-edge LLM model, GPT-4, doesn't show much superiority on this task compared to other LLMs.

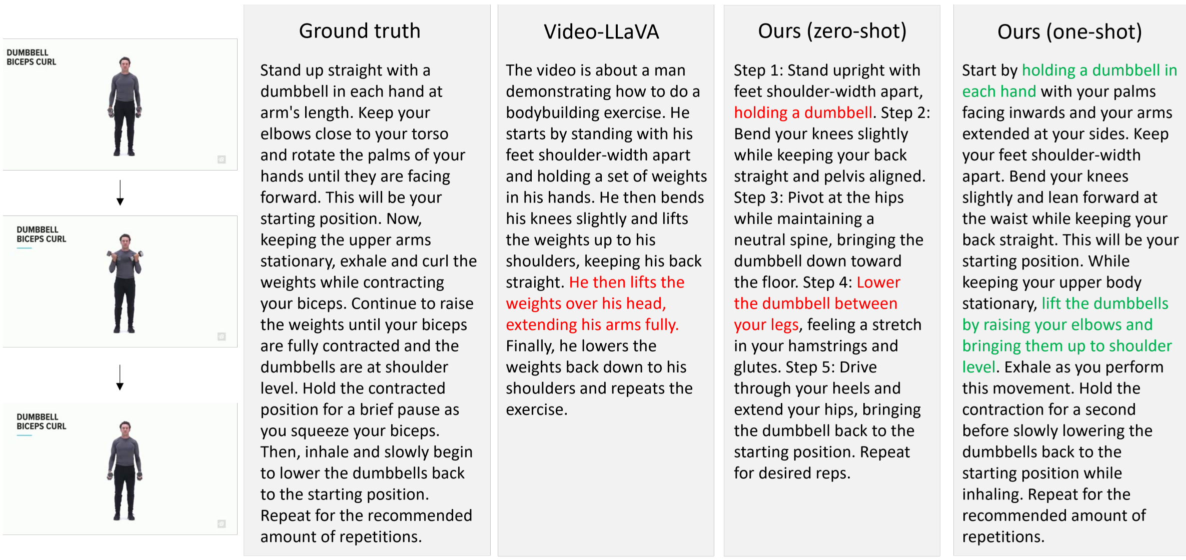


Figure 3: Comparison between text generated by VLM and PoseGPT based on zero-shot and one-shot prompts. Results of PoseGPT are generated by ChatGPT using Tait-Bryan angle representation.

5.4 One-shot performance

Then we add a sample to the prompts and evaluate the one-shot performance of PoseGPT. Results are shown in Table 4. PoseGPT on all backbones obtain better results than zero-shot and ChatGPT performs best. Though the Tait-Bryan angle representation models motion more accurately, it does not contribute to general performance improvement. One possible reason is that LLMs can not fully understand complex rotation data. Another possible reason is that prompts using Tait-Bryan angle representation are significantly longer than those using included angle representation, and longer context makes it more challenging for LLMs to focus on critical angle changes.

5.5 Case study

Figure 3 shows a sample video and caption generated by models. Both the Video-LLaVA results and our zero-shot methods contain factual errors, while one-shot results are of significantly higher quality. Though our prompts do not provide any further information about the equipment except its name, LLM still has some ability to reason the location and quantity of the equipment.

5.6 Frame sampling

We evaluate the changes in FCE-Motion and METEOR scores on ChatGPT with sampling ranging from 5 frames to 10 frames. We use the 16k context version to avoid prompt length overflow. Results are shown in Figure 4. We find that the FCE-

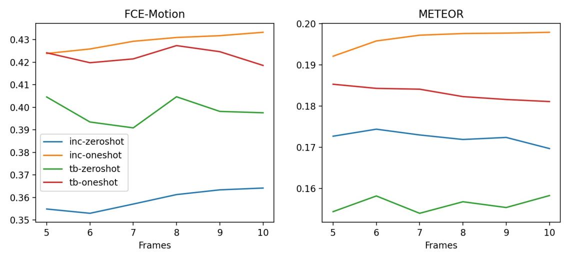


Figure 4: Visualizations of the relationship between evaluation metrics and frame numbers.

Motion score increases slightly with the increase of frame numbers when using included angle representations, indicating a better motion description capability, while the method using the Tait-Bryan angle representation does not show the same trend. This may be because the Tait-Bryan representation's prompt length increases more as the number of frames increases, which has a greater impact on attention.

6 Conclusions

We construct BoFit, a fine-grained fitness training dataset for video captioning. We also propose PoseGPT, 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 PoseGPT outperforms previous VLMs on BoFit on comprehensive metrics.

Limitations

We first propose the fine-grained human motion video captioning task. Since it is difficult to acquire the pairs of videos and their descriptions, the scale of our dataset BoFit is relatively small. In addition, we make use of human posture features as intermediate representations between video and text, which may lose some information in videos. We would like to explore more reasonable intermediate representation to help LLM understand videos.

References

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat-Thalmann. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2272–2281.

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*.

Yu Cheng, Bo Yang, Bo Wang, and Robby T. Tan. 2020. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *ArXiv*, abs/2004.11822.

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.

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.

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.

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 pre-training 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.

Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. 2021. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13137–13146.

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*.

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*.

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.

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*.