LLaSA: Large Multimodal Agent for Human Activity Analysis Through Wearable Sensors

Anonymous EMNLP submission

Abstract

Integrating inertial measurement units (IMUs) with large language models (LLMs) advances multimodal AI by enhancing human activity understanding. We introduce SensorCaps, a dataset of 26,288 IMU-derived activity narrations, and OpenSQA[1], an instruction-following dataset with 257,562 question-answer pairs. Combining LIMU-BERT and Llama, we develop LLaSA, a Large Multimodal Agent capable of interpreting and responding to activity and motion analysis queries. Our evaluation demonstrates LLaSA's effectiveness in activity classification and question answering, highlighting its potential in healthcare, sports science, and human-computer interaction. These contributions advance sensor-aware language models and open new research avenues.

1 Introduction

Developing large language models (LLMs) requires comprehensive question-answering datasets for effective task-specific training and fine-tuning. According to Xie et al., creating such datasets involves curating diverse, high-quality question-answer pairs to guide models in learning context-specific responses and reasoning abilities (Xie et al., 2024). In modalities like audio and video, narrations are used to prepare question-answering datasets. Gong et al. (Gong et al., 2023) integrates the audio modality with LLaMA (Touvron et al., 2023) and uses GPT generated audio narrations (WavCaps) to enhance their dataset.

Building on these ideas, this paper explores integrating inertial measurement units (IMUs) with LLMs to expand their real-world applicability. IMUs, which combine accelerometers, gyroscopes, and magnetometers, provide precise, continuous data on human motion, valuable for applications like healthcare monitoring, sports science, and human-computer interaction. Incorporating IMU data into large multimodal agents (LMAs) can enhance their understanding of the environment, improving decision-making and task execution.

For instance, IMU-equipped LMAs can monitor physical activity and detect anomalies in healthcare, providing timely interventions and personalized recommendations. LIMU-BERT (Xu et al., 2021) demonstrates the effectiveness of integrating IMU data with language models using self-supervised learning on unlabeled IMU data, improving human activity recognition accuracy by over 10%.

To enhance multimodal agent's capabilities, developing comprehensive question-answering datasets is crucial. While previous works developed multimodal datasets to integrate LLM with other modalities (e.g., audio and video) (Gong et al., 2023; Mei et al., 2023), no such dataset exists for IMU data. We present a unique dataset of human activity narrations, capturing IMU events and translating them into detailed descriptions. This dataset is used to generate question-answering datasets, facilitating training LLMs to understand and respond to queries about human activities and motion analysis. We combine Llama and LIMU-BERT to develop a sensor-aware question-answering model, marking a significant advance in multimodal AI.

The contributions of this paper are threefold: (1) introducing and publishing SensorCaps, a novel dataset narrating IMU data into 26,288 human activity, and OpenSQA, an instruction-following dataset with 257,562 question-answer pairs; (2) developing a multimodal model integrating Llama and LIMU-BERT, improving performance in understanding and responding to queries about human activities and motion analysis; and (3) providing a comprehensive performance evaluation in close-ended human activity classification and open-ended question-answering in a new benchmark dataset. These contributions advance multimodal AI and open new research and applications.

[1] Our anonymous code repository and datasets can be found on https://anonymous.4open.science/r/LLaSA/

2 Related Work

Researchers have introduced LMAs for various modalities, including computer vision (Liu et al., 2023; Wang et al., 2023), audio (Huang et al., 2024; Gong et al., 2023), motion animation (Zhang et al., 2024), and health sensors (Kim et al., 2024), demonstrating their versatility and potential in enhancing AI applications. While early LMAs used closed-source LLMs like GPT-3.5 for inference (Xie et al., 2024), recent efforts focus on preparing question-answering datasets and fine-tuning open-source models like Llama (Touvron et al., 2023) with multimodal encoders, such as LLaVA (Liu et al., 2024). These models perform tasks like image and video understanding, video generation or editing, autonomous driving, and game development. However, the application of wearable sensors remains limited, particularly in creating instruction-following datasets.HealthLLM (Kim et al., 2024) integrates LLM with sensor data but focuses solely on close-ended tasks without supporting open-ended question-answering.

Wearable IMU sensors are crucial for understanding human activities. LIMU-BERT (Xu et al., 2021) improves recognition accuracy through self-supervised learning. Penetrative AI (Xu et al., 2024) uses GPT-3.5 and GPT-4 for motion and heartbeat detection but fails to answer questions.

Our work extends research by analyzing fine-tuned LMAs' ability to understand, discuss, and answer questions about human activities using accelerometer and gyroscope data. Narration of signals, as shown by Gong et al. (Gong et al., 2023), is key for creating instruction-following datasets. While narrating wearable sensor data is novel, similar methods in WavCaps (Mei et al., 2023) use GPT for audio event captions. Gong et al. leveraged this to create the OpenAQA dataset. By applying GPT to IMU data, we aim to develop comprehensive human activity based question-answering datasets, laying the foundation for sensor-aware LMAs.

3 Large Language and Sensor Assistant

This section describes the design methodology of the Large Language and Sensor Assistant (LLaSA).

3.1 Foundational Model for Motion Data

As the foundational model for encoding motion data, we use LIMU-BERT, which leverages unlabeled IMU data through self-supervised learning, similar to BERT in Natural Language Processing.

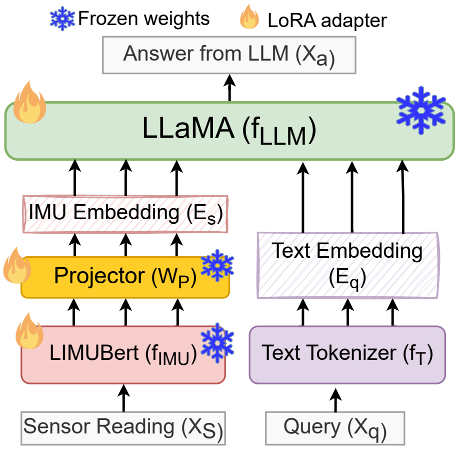

Figure 1: Architecture of the Large Language and Sensor Assistant (LLaSA) model

We first fuse and normalize accelerometer and gyroscope data from the IMU sensors, then apply a trainable positional encoding to fully utilize the order information. The encoder of LIMU-BERT consists of multiple blocks combining multi-headed attention, projection, and feed-forward layers, connected by add & normalization layers. The decoder comprises a projection layer, an activation & normalization layer, and a prediction head. Both the encoder and decoder use the Gaussian Error Linear Unit (GELU) as the activation function.

After training LIMU-BERT with unlabeled data from four activity recognition datasets (HHAR (Stisen et al., 2015), UCI-HAR (Reyes-Ortiz et al., 2016), MotionSense (Malekzadeh et al., 2019), and Shoaib (Shoaib et al., 2014)), we save the encoder and combine it with a language model for further tasks. These datasets provide three-axis accelerometer and gyroscope data for eight activities: "walking," "sitting," "standing," "jogging," "biking," "ascending stairs," "lying," and "descending stairs." These are downsampled to 20 Hz.

LIMU-BERT Training Details. We train LIMU-BERT using the mean square error (MSE) loss and Adam optimizer on a Nvidia RTX 3090Ti GPU for 8 hours.

3.2 Integrating with LLM

Following LLaVA, we merge the projected outputs of the LIMU-BERT encoder with an LLM that encodes the textual inputs and provides textual outputs. We use the 7 billion parameter Vicuna 1.5 model (Chiang et al., 2023), built on the Llama foundational LLM. Figure 1 illustrates the LLaSA model architecture. Here, X_s (sensor readings) goes through LIMU-BERT, f_{IMU}, and Multilayer

Perceptron (MLP) projector,W_P, to become the encoding E_s. X_q (natural language query) is processed by the Llama tokenizer (f_T), producing encoding E_q. These encodings are then input to the LLM agent (f_{LLM}), resulting in the answer X_a. This process is summarized by:

$$\mathrm{X}_a = f_{LLM}\left([f_{IMU}(X_s) * W_P] \cap [f_T(X_q)]\right)$$

Projector Training Details. We pre-train the MLP projector (W_P) with a GELU activation layer using the OpenSQA dataset for one epoch in 16-bit precision on two 24GB GPUs with a batch size of 32 per device and a learning rate of 0.001.

LLaSA Training Details. For parameter-efficient fine-tuning, we use LoRA (Hu et al., 2021) to fine-tune Vicuna-7b-1.5 and the pretrained projector for one epoch with 16 samples per batch, 128 rank, and a 0.0002 learning rate. This follows LLaVA's instruction-following multimodal agent training procedure, enabling the model to interpret sensor data when answering questions.

4 Sensor-Context Aware Instruction Following Dataset Generation

This section outlines the creation of captioning and question-answering datasets.

4.1 SensorCaps: Captioning Dataset

Before preparing the question-answering dataset, we create a sensor captioning or human activity narration dataset with IMU data. It ensures that the instruction-following dataset preparation pipeline can access knowledge about data samples. We use the same datasets mentioned in Section 3.1, subsampling the data to 10 Hz and rounding them to 6 digits to reduce generation costs and prevent the chatbot from focusing on minor details. We then send the sensor data, ground truth labels, and detailed instructions to GPT-3.5-Turbo to narrate the IMU event or human activity. The narration generation pipeline first asks to extract and summarize characteristic features of the sensor data before generating temporally aware captions.

4.2 OpenSQA: Question-Answering Dataset

With SensorCaps, we have four types of information for each sensor data reading: (1) IMU signal values (gyroscope and accelerometer), (2) activity label or summary (e.g., "descending stairs"), (3) summary of characteristic features of the IMU signals, and (4) narration of the IMU event.

We provide these to GPT-3.5-Turbo with detailed prompts to generate ten question-answering pairs that require knowledge to step-by-step analyze the data and context. With the information in SensorCaps, GPT-3.5-Turbo generates a list for building instruction-following training data. Occasionally, GPT-3.5-Turbo fails to generate questions with answers, resulting in the loss of 5,318 pairs, but we still retain 257,562 instruction-following training samples for our LMA, considering human activity analysis from the IMU data context.

5 Evaluation Tasks and Datasets

This section outlines the tasks, datasets, and metrics for close- and open-ended evaluation of LLaSA.

5.1 Close-Ended

Due to the popularity of human activity recognition from IMU data (Chen et al., 2021), we use it as the close-ended zero-shot evaluation task to determine if LLaSA correctly understands IMU data. We evaluate LLaSA on four seen datasets (HHAR, UCI-HAR, MotionSense, and Shoaib) used in its training (Section 3.1) and one unseen dataset (SHL (Gjoreski et al., 2018)) not used to generate OpenSQA. For testing, we use stratified subsets with 100 samples per class from each dataset for balanced representation. For each dataset, we prompt the LLMs with possible activity labels. If the LLM fails to answer with a relevant label, it is classified as "Unclear." Relevant labels match the activity classes from the dataset. We compare the predictions of our proposed model with GPT-3.5 Turbo and a fine-tuned version of GPT-3.5 Turbo, GPT-3.5-T-F (fine-tuned with 5% of the LIMU-BERT training data and corresponding human activity labels). We use precision, recall, and F1-score as metrics to evaluate the model's performance comprehensively.

5.2 Open-Ended

To evaluate LLaSA's ability to answer IMU-related open-ended questions, we develop an open-ended benchmark dataset with 19,440 question-answer pairs covering diverse human activities. We use the PAMAP2 dataset (Roggen et al., 2010), which was not used to train LLaSA. This benchmark includes three categories of questions: 1) scientific depth of knowledge, 2) reasoning behind possible activities, and 3) reliability of sensor readings (e.g., noise effects). For each of the 18 classes, we randomly select two samples from three sensor locations (ankle, chest, hand) for each of the nine subjects, covering

Dataset	LMA	F-1	Precision	Recall
HHAR	LLaSA	0.84	0.88	0.86
	GPT-3.5-T	0.07	0.16	0.10
	GPT-3.5-T-F	0.52	0.54	0.70
MotionSense	LLaSA	0.83	0.85	0.83
	GPT-3.5-T	0.08	0.07	0.13
	GPT-3.5-T-F	0.21	0.27	0.21
Shoaib	LLaSA	0.81	0.84	0.81
	GPT-3.5-T	0.09	0.10	0.12
	GPT-3.5-T-F	0.27	0.45	0.31
UCI	LLaSA	0.72	0.82	0.75
	GPT-3.5-T	0.07	0.35	0.12
	GPT-3.5-T-F	0.28	0.24	0.31
SHL	LLaSA	0.65	0.76	0.71
	GPT-3.5-T	0.15	0.22	0.15
	GPT-3.5-T-F	0.23	0.25	0.32

Table 1: Performance comparison of LLaSA with GPT-3.5-Turbo on human activity recognition tasks

all three categories. These data undergo the same question-answer generation process as SensorCaps, including subsampling and rounding. Using GPT-4o (GPT, 2024), we generate five question-answer pairs per category with appropriate instructions.

6 Experimental Results

6.1 Close-ended

Table 1 shows that in close-ended zero-shot evaluation, GPT-3.5-Turbo cannot associate the signals with the correct classes when the information regarding the ground truth is withheld. On the contrary, LLaSA takes embedding instead of the IMU signal strings like GPT-3.5-Turbo and has 12, 10.4, 9, and 10.3 times higher F1-Score on HHAR, MotionSense, Shoaib, and UCI data, respectively. Although fine-tuned GPT-3.5 performed better than non-finetuned GPT-3.5-Turbo, LLaSA outperforms it with an average of 2.8 times higher F1-score. For the unseen dataset during training (SHL), we observe that LLaSA achieves a 65% F1-score whereas GPT-3.5-Turbo and fine-tuned GPT-3.5-Turbo have an F1-score of 15% and 23%, respectively.

6.2 Open-ended

For evaluation, we ask GPT-4o to assess our predicted answers against the ground truth activity label, sensor location, and the benchmark dataset's standard answer generated by GPT-4o. This evaluation involves both human observation and GPT-4o's summary of LLaSA's strengths and weaknesses.

Table 2 shows example question-answering with LLaSA, which can correlate IMU data with various activities, such as the stance and swing phases of stair climbing, providing relevant interpretations. Vicuna-13b, on the contrary, provides general answers instead of analyzing the data. Though LLaSA may misclassify, discuss unrelated sensors, and provide general answers without specific details, it can still offer insights from scientific, narrative, and data reliability perspectives.

User	What is this person doing? [ascending stairs]
LLaSA	The accelerometer data shows a consistent increase in the z-axis values, which aligns with the upward movement typically associated with climbing stairs.
Vicuna-13b	[omitted]... It is difficult to determine the specific activity or context of the movements without additional information. However, the data could be used to analyze the person's gait, balance, or other movements, or to track the device's movement over time.
User	What does the combination of gyroscope and accelerometer data reveal about the terrain on which the running is likely occurring? [running]
LLaSA	The gyroscope data showing rapid changes in orientation and the accelerometer data displaying high acceleration values suggest that the individual is likely running on uneven terrain with obstacles or inclines. The combination of these sensor readings indicates a challenging environment that requires quick adjustments in movement and speed.
Vicuna-13b	[omitted]... Overall, the combination of gyroscope and accelerometer data can provide valuable information about the terrain on which the running is occurring, ... [omitted]
User	How can the gyroscope data be used to identify potential fatigue or inconsistencies in the hand movement while ironing? [ironing]
LLaSA	[omitted] ... by analyzing the smoothness and consistency of the rotational movements. Sudden spikes or irregular patterns in the gyroscope data may indicate fatigue or inconsistencies in the hand movements, suggesting potential challenges in maintaining a steady and controlled ironing motion.
Vicuna-13b	[omitted] ... One approach is to calculate the standard deviation of the angular velocity over time to identify if the iron is being moved with consistent force or if there are periods of increased or decreased force that could indicate fatigue. Another approach is to analyze the orientation of the iron over time to identify if there are any patterns or deviations from a straight line that could indicate inconsistent movement or fatigue. ... [omitted]

Table 2: Sample question-answering with LLaSA

7 Discussion and Conclusion

This paper introduces a novel LMA for IMU data that links human activity narratives with sensor readings. We develop SensorCaps, OpenSQA, and IMU-aware question-answering benchmarks, advancing multimodal NLP research. Our experiments show that the LLaSA model effectively associates IMU data with activity phases and provides insightful responses. While it occasionally misclassifies activities and struggles with mathematical problems, these issues highlight areas for future improvement. The potential applications in healthcare, sports science, and human-computer interactions underscore the value of our work.

8 Limitations

This section outlines the limitations of our proposed LMA, LLaSA.

8.1 Generating Detailed and Specific Answers

Our open-ended question-answering model sometimes produces overly generalized answers. This could be mitigated by creating datasets with more detailed and specific answers. In the future, we plan to focus on generating diverse question-answer categories to ensure the LMA can respond with specialized details instead of relying on a limited knowledge.

8.2 Mathematical problem solving with sensor-aware LMAs

We generate a fourth category of questions for the benchmark dataset in Section 5.2 to solve small mathematical problems. Our experiments revealed that sensor embeddings alone are insufficient for solving mathematical problems involving sensor data. GPT-4o assessment summaries also include this as a weakness of LLaSA. In our future work, we will consider appending numerical sensor readings to queries. We plan to investigate whether enhancing mathematical problem-solving capabilities can improve overall understanding in other instruction-following or question-answering tasks.

8.3 Open-ended question-answering evaluation

While using LLMs like GPT-4o to summarize strengths and weaknesses provides useful insights, developing metrics or scoring systems based on these summaries could help compare generative models. In the future, we will explore metrics that assess the accuracy of references to physical activities and sensors, incorporating human expert assessments to validate these metrics. The alignment between a human scorer and an LLM scorer can verify the quality of such metrics.

8.4 Hyper-parameter search

This work does not explore optimal hyper-parameter search for training LMAs. It will be worth investigating hyper-parameter optimization to enhance training procedures and performance.

8.5 Full fine-tuning and bigger models

We used a relatively lightweight LLM (7 billion parameters) with parameter-efficient fine-tuning (PEFT). Our future work will explore larger models (e.g., over 30 billion parameters) without PEFT. Such models, combined with tuned hyper-parameters, might better handle complex data.

8.6 Ethical considerations for future research

Until further research ensures the safety of using LMAs in real-life activities, we advise against working with human subjects. Although our study does not directly address the potential risks, incorrect answers and hallucinations from LMAs could misguide and endanger users, especially those relying on wearable technologies for health. Therefore, our future research will focus on improving our understanding of the safety of LMAs in wearable and environmental sensor applications.

Ethics Statement

The IMU data used in this paper are publicly available online and were collected and distributed by third parties with consent and IRB or ethics committee approvals where applicable. We did not use any private IMU data to train the model. The proposed model has the potential to benefit individuals by providing detailed information about activities. However, while this model aims to classify and detail activities accurately, it may occasionally provide incorrect answers, which could result in dangerous outcomes if misused. Therefore, it is crucial for researchers, developers, and users to employ this technology responsibly, ensuring its application aligns with ethical considerations and avoids potential misuse. At its current stage, the models and methods presented in this paper are intended solely for research purposes and should not be used outside research circles or provided to human subjects or consumers.

References

2024. Hello gpt-4o. *OpenAI*.

Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54(4):1–40.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hristijan Gjoreski, Mathias Ciliberto, Lin Wang, Francisco Javier Ordonez Morales, Sami Mekki, Stefan Valentin, and Daniel Roggen. 2018. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. *IEEE Access*, 6:42592–42604.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

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*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*, pages 49–58.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing*, 171:754–767.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pages 233–240. IEEE.

Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14(6):10146–10176.

Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pages 127–140.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, pages 1–7.

Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 220–233.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2024. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7368–7376.