RehabGPT: A MaaS based solution to large model building for digital rehabilitation

Hongsheng Wang[1,2]
[1]Zhejiang Lab,[2]Zhejiang University
wanghongsheng@zhejianglab.com

Zhangnan Zhong[3]
[3]Shenzhen University
2110246043@email.szu.edu.cn

Linwei Dai[4]
[4]Xidian University
21171213925@stu.xidian.edu.cn

Fei Wu[2]
[2]Zhejiang University
wufei@zju.edu.cn

Xiao Ma[1]
[1]Zhejiang Lab
mx@zhejianglab.com

Junxiao Xue[1]
[1]Zhejiang Lab
xuejx@zhejianglab.com

Phuminh Lam[1], Smirnov Pavel[1]
[1]Zhejiang Lab
Pavel@zhejianglab.com

Qing Zhang[1]
[1]ZhejiangLab
qing.zhang@zhejianglab.com

Feng Lin[1,*]
[1]Zhejiang Lab
asflin@zhejianglab.com

Abstract — **Digital rehabilitation plays a crucial role in the treatment of chronic diseases, as it enables the assessment of disease grades and the recommendation of treatment measures. In this paper, we propose a generative pre-trained transformer towards rehabilitation (RehabGPT) via a model-as-a-service (MaaS) solution to facilitate foundation model building for digital rehabilitation on Alibaba's ModelScope platform. It offers scalable computational resources needed for pre-trained large models. It also provides tools for multi-modal feature extraction, 3D human mesh reconstruction and analysis of video sequences. RehabGPT automates various aspects of the model development workflow, such as hyperparameter tuning and architecture selection, making it easier to achieve the desired results in rehabilitation tasks.**

Keywords—MaaS, RehabGPT, Foundation models, Digital rehabilitation

I. Introduction

ModelScope is an artificial intelligence model development platform launched by Alibaba DAMO Academy for foundation models development and deployment. It provides a variety of foundation models, based on which domain-specific models can be generated. ModelScope includes foundation models such as ActionCLIP[1] and SAM[2], which are capable of generating images from text.

II. Methodology

During the process of 3D human mesh reconstruction from video sequences, we propose SAM-Track to emphasize the fusion of features among frames. This approach not only leverages SAM's inherent powerful semantic feature extraction capabilities but also complements the information from video sequences. To improve the global interactions between joints and vertices, we employ transformer encoder-decoder architecture to directly reconstruct 3D human mesh from video sequences, then analyze the reconstructed sequences and provide digital rehabilitation suggestions according to the patient's performance. During the modeling process for rehabilitation video analysis, to further amplify the model's performance and accuracy, we propose a foundation model with contrastive learning, which was modeled as a text-to-video matching task, enhancing video representation through more linguistic semantic supervision. This method emphasizes utilizing semantic relations and similarity measures among multi-modal data, ensuring scientifically accurate guidance for patients undergoing digital rehabilitation across various disease grades.

III. Conclusion

In summary, this study presents a approach to leverage the capabilities from ModelScope's text-to-image model generation. Based on the reconstructed 3D human mesh sequences, the approach seeks to offer scientific suggestions for digital rehabilitation. To ensure the precision and reliability of performance, future research will focus on the development and integration of foundation models tailored for specific domains.

Acknowledgment

This work is partially supported by Zhejiang Lab & Pujiang Lab, China (K2023KA1BB01), with their grants 111000-BB2201, 111000-BB2301 & 111000-PI2201.

References

[1] Mengmeng W, Jiazheng X, Yong L. ActionCLIP: A New Paradigm for Video Action Recognition[J], CoRR, 2021, abs/2109.08472

[2] Alexander K, Eric M, Nikhila R, Hanzi M, Chloe R, Laura G, Tete X, Spencer W, Alexander C B, Wan-Yen L, Piotr D, Ross G, et al. Segment Anything[J], CoRR, 2023, abs/2304.02643