SIPENG ZHENG

zhengsipeng27@gmail.com zhengsipeng.github.io Google Scholar +86 15905058181

I am a partner at BeingBeyond, a startup dedicated to advancing foundation models for embodied AI, where I collaborate closely Prof. Zongqing Lu. Now I am leading the Embodied Multimodal Pretraining team in BeingBeyond, with projects including Being-H, Being-M and Being-VL series. Before that, I was a research scientist at the Beijing Academy of Artificial Intelligence (BAAI). I received both my PhD and bachelor degrees from Renmin University of China in 2023 and 2018, respectively, under the supervision of Prof. Qin Jin. My research interests focus on human behavior and motion understanding, vision-language pretraining, and embodied artificial intelligence. Currently, I am working on developing general-purpose humanoid robots.

EDUCATION

Renmin University of China *2018.09 - 2023.06*
Ph.D. in Computer Science Advisor: Qin Jin

Renmin University of China *2014.09 - 2018.06*
B.S. in Computer Science Advisor: Qin Jin

KEY PROJECTS

Being-H0 *2025.07*
This groundbreaking vision-language-action model is the first to be pretrained on large-scale human videos showcasing hand movements. It demonstrates significant advancements in performance across various dexterous tasks, notably reducing the demand for extensive training data. We are actively developing its next iteration, aiming to enhance its capabilities and generalization.

Being-M0.5 *2025.08*
Our large-scale, sota-performance motion generation foundation model with real-time controllability, trained by over 1 million self-collected motion data based on our self-built data curation pipeline. **Being-M1** is on the way.

Being-VL-0.5 *2025.07*
Our large-scale vision-language model with a novel image tokenizer that bridges this gap by applying the principle of Byte-Pair Encoding (BPE) to visual data.

Being-0 *2025.2*
This is a hierarchical agent framework that integrates an Foundation Model (FM) with a modular skill library. The foundation model is utilized to handle high-level cognitive tasks such as instruction understanding, task planning, and reasoning, while the skill library provides stable locomotion and dexterous manipulation for low-level control.

WORK EXPERIENCE

BeingBeyond **Beijing, China**
General partner, Research scientist *2025.05 - Present*

- Lead research of Embodied Multimodal Pretraining team, with projects including **Being-M** (human motion generation), **Being-H** (hand pose generation) and **Being-VL** (multimodal model).
- Scaled pretraining to billion-level image-text and million-level video datasets.
- Hands-on deployment with humanoid robots: Unitree G1/H1/H1-2, Fourier GR1.

Beijing Academy of Artificial Intelligence **Beijing, China**
Research scientist *2023.07 - 2025.05*

Beijing Academy of Artificial Intelligence **Beijing, China**
Research intern *2021.09 - 2022.02*

- Multi-lingual language-vision-audio pre-training.

Microsoft Research Asia **Beijing, China**
Research intern *2022.04 - 2022.10*

- Temporal sentence grounding for long-term videos.

AWARDS

- Ranked 1st GemBench Challenge at CVPR 2025 Workshop GRAIL. 2025
- Ranked 3th in Facebook CVPR 2022 Ego4D Natural Language Query Challenge. 2022
- Ranked 3th, NIST TRECVID 2021 Ad-hoc Video Search (AVS) Challenge. 2021
- Ranked 2nd in CVPR 2021 HOMAGE Scene-graph Generation Challenge. 2021
- Ranked 2nd in ACM MM 2020 Video Relationship Understanding Grand Challenge. 2020
- Ranked 2nd in ACM MM 2019 Video Relationship Understanding Grand Challenge. 2019
- National Scholarship for Ph.D Students (Top 5%) 2022
- Best Method Prize in ACM MM 2019 Grand Challenge 2019
- First Class Scholarship for Ph.D Students (Top 10%) 2018-2021
- First Prize in National University Mathematical Modeling Competition. 2015

PUBLICATION

*** denotes equal contribution**

1. Hao* Luo, Yicheng* Feng, Wanpeng* Zhang, **Zheng, Sipeng***, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-H0: Vision-Language-Action pretraining from large-scale human videos. *arxiv*, 2025

2. Junpeng Yue, Zepeng Wang, Yuxuan Wang, Weishuai Zeng, Jiangxing Wang, Xinrun Xu, Yu Zhang, **Sipeng Zheng**, Ziluo Ding, and Zongqing Lu. Rl from physical feedback: Aligning large motion model with robot whole body control. *arxiv*, Neurips 2025 Submission

3. Hao Luo, Zihao Yue, Wanpeng Zhang, Yicheng Feng, **Sipeng Zheng**, Deheng Ye, and Zongqing Lu. EgoCatch: Enhancing egocentric video understanding for multimodal llms. *arxiv*, Neurips 2025 Submission

4. Boshen Xu, Yuting Mei, Xinbi Liu, **Sipeng Zheng**, and Qin Jin. EgoDTM: Towards 3d-aware egocentric video-language pretraining. *arxiv*, Neurips 2025 Submission

5. Jiazheng Liu, Börje F. Karlsson, **Sipeng Zheng**, and Zongqing Lu. Taking notes brings focus? towards multi-turn multimodal dialogue learning. *EMNLP*, 2025

6. Yuting Mei, Ye Wang, **Sipeng Zheng**, and Qin Jin. Integrating path planning and adaptive locomotion for mobile quadruped robots with large multimodal models. *arxiv*, 2025

7. Wanpeng Zhang, Yicheng Feng, Hao Luo, Yijiang Li, Zihao Yue, **Sipeng Zheng**, and Zongqing Lu. Unified multimodal understanding via byte-pair visual encoding. *ICCV*, 2025

8. Bin Cao*, **Sipeng Zheng***, Ye Wang, Lujie Xia, Qianshan Wei, Qin Jin, Jing Liu, and Zongqing Lu. Motionctrl: A real-time controllable vision-language-motion model. *ICCV*, 2025

9. Yicheng Feng, Yijiang Li, Wanpeng Zhang, **Sipeng Zheng**, and Zongqing Lu. Videoorion: Tokenizing object dynamics in videos. *ICCV*, 2025

10. Ye Wang*, **Sipeng Zheng***, Bin Cao, Qianshan Wei, Weishuai Zeng, and Zongqing Lu. Scaling large motion models with million-level human motions. *ICML*, 2025

11. Wanpeng Zhang, Zilong Xie, Yicheng Feng, Yijiang Li, Xingrun Xing, **Sipeng Zheng**, and Zongqing Lu. From pixels to tokens: Byte-pair encoding on quantized visual modalities. *ICLR*, 2025

12. Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, **Sipeng Zheng**, and Qin Jin. EgoNCE++: Do egocentric video-language models really understand hand-object interactions? *ICLR*, 2025

13. Boshen Xu, **Sipeng Zheng**, and Qin Jin. SPAFormer: Sequential 3d part assembly with transformers. *3DV*, 2025

14. **Sipeng Zheng**, Bohan Zhou, Yicheng Feng, Ye Wang, and Zongqing Lu. UniCode: Learning a unified codebook for multimodal large language models. *ECCV*, 2024

15. **Sipeng Zheng**, Jiazheng Liu, Yicheng Feng, and Zongqing Lu. Steve-Eye: Equipped llm-based embodied agents with visual perception in open worlds. *ICLR*, 2024

16. Yicheng Feng, Yuxuan Wang, Jiazheng Liu, **Sipeng Zheng**, and Zongqing Lu. LlaMA Rider: Spurring large language models to explore the open worlds. *NAACL*, 2024

17. Qi Zhang, **Sipeng Zheng**, and Qin Jin. No-frills temporal video grounding: Multi-scale neighboring attention and zoom-in boundary detection. In *AAAI*, 2023

18. **Sipeng Zheng**, Boshen Xu, and Qin Jin. Open-category human-object interaction pre-training via language modeling framework. In *CVPR*, 2023

19. Boshen Xu, **Sipeng Zheng**, and Qin Jin. POV: Prompt-oriented view-agnostic learning for egocentric hand-object interaction in the multi-view world. *ACM MM*, 2023

20. Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, **Sipeng Zheng**, and Qin Jin. Accommodating audio modality in clip for multimodal processing. *AAAI*, 2023

21. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *ECCV*, 2022

22. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. VRDFormer: End-to-end video visual relation detection with transformers. In *CVPR*, 2022 (Oral)

23. **Sipeng Zheng**, Qi Zhang, and Qin Jin. Exploring anchor-based detection for ego4d natural language query. In *CVPR Ego4D workshop*, 2022

24. Bei Liu, **Sipeng Zheng**, Jianlong Fu, and Wen-Huang Cheng. Anchor-based detection for natural language localization in ego-centric videos. In *IEEC*, 2022

25. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Skeleton-based interactive graph network for human object interaction detection. In *ICME*, 2020

26. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Visual relation detection with multi-level attention. In *ACM MM*, 2019

27. **Sipeng Zheng**, Xiangyu Chen, Shizhe Chen, and Qin Jin. Relation understanding in videos. In *ACM MM*, 2019

SERVICES

- Conference Reviewer for CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, AAAI, ACM MM.

- Journal Reviewer for IJCV, TCSVT, TMM, JATS.