SIPENG ZHENG

✉ zhengsipeng27@gmail.com ⌂ zhengsipeng.github.io Sipeng Zheng ☎ +86 15905058181

I am a research scientist at the Beijing Academy of Artificial Intelligence (BAAI), working with Prof. Zongqing Lu. 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 learning, human motion understanding, vision-and-language pretraining, and embodied artificial intelligence. Currently, I am working on developing general-purpose embodied agents, including game agents and humanoid robots.

EDUCATION

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

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

WORK EXPERIENCE

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

- Experience in human motion understanding and generation, as well as human-object interaction.
- Skilled in pretraining, data curation, and deploying large multimodal models on embodied agents, including humanoid robots, human-computer interaction systems, and open-world agents (e.g., Minecraft).
- Experience in large-scale pretraining on billion-level image-text and million-level video data.
- Hands-on experience with humanoid robots, including models such as H1, H1-2, GR1, G1, Adam, GO2, and others.

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

- ⋆ National Scholarship for Ph.D Students. 2022
- ⋆ 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 4th 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
- ⋆ Best Method Prize in ACM MM 2019 Grand Challenge 2019
- ⋆ First Class Scholarship for Ph.D Students. 2018-2021
- ⋆ First Prize in National University Mathematical Modeling Competition. 2015

PROFESSIONAL ACTIVITIES

- ⋆ Conference Reviewer for CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, AAAI, ACM MM.
- ⋆ Journal Reviewer for IJCV, TCSVT, TMM.

PUBLICATION

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

2. 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 Submission

3. 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 Submission

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

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

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

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

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

9. 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

10. 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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