

SIPENG ZHENG

✉ zhengsipeng27@gmail.com 🏠 [zhengsipeng.github.io](https://github.com/zhengsipeng) 📄 [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 Ph.D. in Computer Science	<i>2018.09 - 2023.06</i> Advisor: Qin Jin
Renmin University of China B.S. in Computer Science	<i>2014.09 - 2018.06</i> Advisor: Qin Jin

KEY PROJECTS

Being-H0 & H0.5 Robots do not just look different. They also act through different physical control languages: different kinematics, sensors, action conventions, and timing. Being-H0.5 is our attempt to make one Vision-Language-Action model travel across those differences without turning into a brittle collection of per-robot hacks. We are actively developing its next iteration, aiming to enhance its capabilities and generalization.	<i>2026.01</i>
Being-M0.5 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.	<i>2025.08</i>
Being-VL-0.5 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.	<i>2025.07</i>
Being-0 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.	<i>2025.2</i>

WORK EXPERIENCE

BeingBeyond <i>Partner, Research Scientist</i>	Beijing, China <i>2025.05 - Present</i>
<ul style="list-style-type: none">◦ Lead research of Embodied Multimodal Pretraining team, with projects including Being-H (hand pose generation), Being-M (human motion 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 <i>Research Scientist</i>	Beijing, China <i>2023.07 - 2025.05</i>

Beijing Academy of Artificial Intelligence
Research Intern

Beijing, China
2022.12 - 2023.06

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

Microsoft Research Asia
Research Intern

Beijing, China
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, † denotes project lead**

- Hao Luo^{*}, Yi Wang^{*}, Wanpeng Zhang^{*}, **Sipeng Zheng^{*†}**, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-H0.5: Scaling Human-Centric Robot Learning for Cross-Embodiment Generalization. *arxiv*, 2026
- Hao Luo^{*}, Yicheng Feng^{*}, Wanpeng Zhang^{*}, **Sipeng Zheng^{*}**, Ye Wang^{*}, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Haiweng Xu, Qin Jin, and Zongqing Lu. Being-H0: Vision-Language-Action Pre-training from Large-Scale Human Videos. *arxiv*, 2026
- Wanpeng Zhang, Ye Wang, Hao Luo, Haoqi Yuan, Yicheng Feng, **Sipeng Zheng**, Qin Jin, and Zongqing Lu. DiG-Flow: Discrepancy-Guided Flow Matching for Robust VLA Models. *arxiv*, 2026
- 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*, 2026
- Boyuan Li, **Sipeng Zheng**, Bin Cao, Ruihua Song, and Zongqing Lu. Robust Motion Generation using Part-level Reliable Data from Videos. *arxiv*, 2026
- Bin Cao, **Sipeng Zheng**, Hao Luo, Boyuan Li, Jing Liu, and Zongqing Lu. OpenT2M: No-frill Motion Generation with Open-source, Large-scale, High-quality Data. *CVPR*, 2026
- Yicheng Feng, Wanpeng Zhang, Ye Wang, Hao Luo, Haoqi Yuan, **Sipeng Zheng**, and Zongqing Lu. Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos. *CVPR*, 2026
- Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, **Sipeng Zheng**, and Zongqing Lu. Predictive Embedding as Latent Action: Towards VLA Pretraining in the Wild. *CVPR*, 2026

9. Hao Luo, Zihao Yue, Wanpeng Zhang, Yicheng Feng, **Sipeng Zheng**, Deheng Ye, and Zongqing Lu. OpenMMEgo: Enhancing Egocentric Understanding for LMMs with Open Weights and Data. *NeurIPS*, 2025
10. Boshen Xu, Yuting Mei, Xinbi Liu, **Sipeng Zheng**, and Qin Jin. EgoDTM: Towards 3D-Aware Egocentric Video-Language Pretraining. *NeurIPS*, 2025
11. Jiazheng Liu, Börje F. Karlsson, **Sipeng Zheng**, and Zongqing Lu. Taking Notes Brings Focus? Towards Multi-Turn Multimodal Dialogue Learning. *EMNLP*, 2025
12. 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
13. 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 (**Highlight**)
14. 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
15. Yicheng Feng, Yijiang Li, Wanpeng Zhang, **Sipeng Zheng**, and Zongqing Lu. VideoOrion: Tokenizing Object Dynamics in Videos. *ICCV*, 2025
16. Ye Wang^{*}, **Sipeng Zheng**^{*}, Bin Cao, Qianshan Wei, Weishuai Zeng, and Zongqing Lu. Being-M0: Scaling Large Motion Models with Million-Level Human Motions. *ICML*, 2025
17. 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
18. 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
19. Boshen Xu, **Sipeng Zheng**, and Qin Jin. SPAFormer: Sequential 3D Part Assembly with Transformers. *3DV*, 2025
20. **Sipeng Zheng**, Bohan Zhou, Yicheng Feng, Ye Wang, and Zongqing Lu. UniCode: Learning a Unified Codebook for Multimodal Large Language Models. *ECCV*, 2024
21. **Sipeng Zheng**, Jiazheng Liu, Yicheng Feng, and Zongqing Lu. Steve-Eye: Equipped LLM-based Embodied Agents with Visual Perception in Open Worlds. *ICLR*, 2024 (**Spotlight 5.02%**)
22. Yicheng Feng, Yuxuan Wang, Jiazheng Liu, **Sipeng Zheng**, and Zongqing Lu. LLaMA Rider: Spurring Large Language Models to Explore the Open Worlds. *NAACL*, 2024
23. Qi Zhang, **Sipeng Zheng**, and Qin Jin. No-frills Temporal Video Grounding: Multi-Scale Neighboring Attention and Zoom-in Boundary Detection. In *AAAI*, 2023
24. **Sipeng Zheng**, Boshen Xu, and Qin Jin. Open-Category Human-Object Interaction Pre-training via Language Modeling Framework. In *CVPR*, 2023
25. 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
26. Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, **Sipeng Zheng**, and Qin Jin. Accommodating audio modality in CLIP for multimodal processing. *AAAI*, 2023
27. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Few-shot Action Recognition with Hierarchical Matching and Contrastive Learning. In *ECCV*, 2022

28. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. VRDFormer: End-to-End Video Visual Relation Detection With Transformers. In *CVPR*, 2022 (**Oral 4.14%**)
29. **Sipeng Zheng**, Qi Zhang, and Qin Jin. Exploring Anchor-based Detection for Ego4D Natural Language Query. In *CVPR Ego4D workshop*, 2022
30. Bei Liu, **Sipeng Zheng**, Jianlong Fu, and Wen-Huang Cheng. Anchor-Based Detection for Natural Language Localization in Ego-Centric Videos. In *IEEC*, 2022
31. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Skeleton-Based Interactive Graph Network For Human Object Interaction Detection. In *ICME*, 2020
32. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Visual Relation Detection with Multi-Level Attention. In *ACM MM*, 2019
33. **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.