Siteng Huang

siteng.huang@gmail.com · kyonhuang.top · Google Scholar (**1000+** citations) · Github (**5k+** stars)

Work Experience

DAMO Academy, Alibaba Group — Hangzhou, China
Algorithm Expert — Sep. 2024 – Present

- Responsible for the development of embodied multimodal large models

TongYi Lab & DAMO Academy, Alibaba Group — Hangzhou, China
Research Intern — Mar. 2022 – July. 2024

- Fundamental Visual Intelligence Team for Tongyi Wanxiang (WanX)
- Directors: Gong Biao, Yu Liu, and Deli Zhao

Education

Zhejiang University — Hangzhou, China
Ph.D. - Computer Science — Sep. 2019 – Jun. 2024

- Thesis: Model Transfer for Multimodal Understanding and Generation

Westlake University — Hangzhou, China
Visiting Student & Joint Program — Oct. 2018 – Jun. 2024

- Affiliated with Machine Intelligence Laboratory (MiLAB), advisor: Prof. Donglin Wang

Wuhan University — Wuhan, China
Bachelor of Engineering - Software Engineering — Sep. 2015 – Jun. 2019

- GPA: 3.7/4.0, Rank: 3/244

Peer-reviewed Journal Publications (∗: equal contribution, ✉: corresponding author.)

[J1] **M2IST: Multi-Modal Interactive Side-Tuning for Memory-efficient Referring Expression Comprehension**
Xuyang Liu, Ting Liu, **Siteng Huang**✉, Yue Hu, Quanjun Yin, Donglin Wang, Honggang Chen✉
IEEE Transactions on Circuits and Systems for Video Technology (TCSVT 2025)

Peer-reviewed Conference Publications (∗: equal contribution, ✉: corresponding author.)

[C1] **CARP: Visuomotor Policy Learning via Coarse-to-Fine Autoregressive Prediction**
Zhefei Gong, Pengxiang Ding, Shangke Lyu, **Siteng Huang**, Mingyang Sun, Wei Zhao, Zhaoxin Fan, Donglin Wang
International Conference on Computer Vision 2025 (ICCV 2025)

[C2] **QUART-Online: Latency-Free Large Multimodal Language Model for Quadruped Robot Learning**
Xinyang Tong, Pengxiang Ding, Donglin Wang, Wenjie Zhang, Can Cui, Mingyang Sun, Yiguo Fan, Han Zhao, Hongyin Zhang, Yonghao Dang, **Siteng Huang**, Shangke Lyu
The 2025 IEEE International Conference on Robotics and Automation (ICRA 2025)

[C3] **Accelerating Diffusion Transformers with Token-wise Feature Caching**
Chang Zou*, Xuyang Liu*, Ting Liu, **Siteng Huang**, Linfeng Zhang
The 13th International Conference on Learning Representations (ICLR 2025)

[C4] **Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference**
Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, **Siteng Huang**, Donglin Wang
The 39th AAAI Conference on Artificial Intelligence (AAAI 2025)

[C5] **ProFD: Prompt-Guided Feature Disentangling for Occluded Person Re-Identification**
Can Cui*, **Siteng Huang***, Wenxuan Song, Pengxiang Ding, Zhang Min, Donglin Wang
ACM Multimedia 2024 (ACMMM 2024)

[C6] **PiTe: Pixel-Temporal Alignment for Large Video-Language Model**
Yang Liu, Pengxiang Ding, **Siteng Huang**, Min Zhang, Han Zhao, Donglin Wang
European Conference on Computer Vision 2024 (ECCV 2024)

[C7] **QUAR-VLA: Vision-Language-Action Model for Quadruped Robots**
Pengxiang Ding, Han Zhao, Wenxuan Song, Wenjie Zhang, Min Zhang, **Siteng Huang**, Ningxi Yang, *et al.*
European Conference on Computer Vision 2024 (ECCV 2024)

[C8] **Learning Disentangled Identifiers for Action-Customized Text-to-Image Generation**
Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, Donglin Wang
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (CVPR 2024)

[C9] **Troika: Multi-Path Cross-Modal Traction for Compositional Zero-Shot Learning**
Siteng Huang, Biao Gong, Yutong Feng, Yiliang Lv, Donglin Wang
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (CVPR 2024)

[C10] **Check, Locate, Rectify: A Training-Free Layout Calibration System for Text-to-Image Generation**
Biao Gong*, **Siteng Huang***, Yutong Feng, Shiwei Zhang, Yuyuan Li, Yu Liu
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (CVPR 2024)

[C11] **DARA: Domain- and Relation-aware Adapters Make Parameter-efficient Tuning for Visual Grounding**
Ting Liu, Xuyang Liu, **Siteng Huang**, Honggang Chen, Quanjun Yin, Long Qin, Donglin Wang, Yue Hu
IEEE Conference on Multimedia Expo 2024 (ICME 2024)

[C12] **VGDiffZero: Text-to-image Diffusion Models Can Be Zero-shot Visual Grounders**
Xuyang Liu*, **Siteng Huang***, Yachen Kang, Honggang Chen, Donglin Wang
IEEE International Conference on Acoustics, Speech and Signal Processing 2024 (ICASSP 2024)

[C13] **Prompt-based Distribution Alignment for Unsupervised Domain Adaptation**
Shuanghao Bai, Min Zhang, Wanqi Zhou, **Siteng Huang**, Zhirong Luan, Donglin Wang, Badong Chen
The 38th AAAI Conference on Artificial Intelligence (AAAI 2024)

[C14] **VoP: Text-Video Co-operative Prompt Tuning for Cross-Modal Retrieval**
Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, Donglin Wang
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR 2023)

[C15] **Reference-Limited Compositional Zero-Shot Learning**
Siteng Huang, Qiyao Wei, Donglin Wang
ACM International Conference on Multimedia Retrieval 2023 (ICMR 2023)

[C16] **Tree Structure-Aware Few-Shot Image Classification via Hierarchical Aggregation**
Min Zhang, **Siteng Huang**, Wenbin Li, Donglin Wang
European Conference on Computer Vision 2022 (ECCV 2022)

[C17] **Domain Generalized Few-shot Image Classification via Meta Regularization Network**
Min Zhang, **Siteng Huang**, Donglin Wang
IEEE International Conference on Acoustics, Speech and Signal Processing 2022 (ICASSP 2022)

[C18] **HINFShot: A Challenge Dataset for Few-Shot Node Classification in Heterogeneous Information Network**
Zifeng Zhuang, Xintao Xiang, **Siteng Huang**, Donglin Wang
ACM International Conference on Multimedia Retrieval 2021 (ICMR 2021)

[C19] **Pareto Self-Supervised Training for Few-Shot Learning**
Zhengyu Chen, Jixie Ge, Heshen Zhan, **Siteng Huang**, Donglin Wang
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021 (CVPR 2021)

[C20] **Attributes-Guided and Pure-Visual Attention Alignment for Few-Shot Recognition**
Siteng Huang, Min Zhang, Yachen Kang, Donglin Wang
The 35th AAAI Conference on Artificial Intelligence (AAAI 2021)

[C21] **DSANet: Dual Self-Attention Network for Multivariate Time Series Forecasting**
Siteng Huang, Donglin Wang, Xuehan Wu, Ao Tang
The 28th ACM International Conference on Information and Knowledge Management (CIKM 2019)

Preprints & Under Review (Only including publicly available work.)

[P1] **WorldVLA: Towards Autoregressive Action World Model**
Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, **Siteng Huang**, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, Hao Chen

[P2] **VARD: Efficient and Dense Fine-Tuning for Diffusion Models with Value-based RL**
Fengyuan Dai*, Zifeng Zhuang*, Yufei Huang, **Siteng Huang**, Bangyan Liao, Donglin Wang, Fajie Yuan

[P3] **SSR: Enhancing Depth Perception in Vision-Language Models via Rationale-Guided Spatial Reasoning**
Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, **Siteng Huang**, Donglin Wang

[P4] **OpenHelix: A Short Survey, Empirical Analysis, and Open-Source Dual-System VLA Model for Robotic Manipulation**
Can Cui, Pengxiang Ding, Wenxuan Song, Shuanghao Bai, Xinyang Tong, Zirui Ge, Runze Suo, Wanqi Zhou, Yang Liu, Bofang Jia, Han Zhao, **Siteng Huang**, Donglin Wang

[P5] **Unicorn: Text-Only Data Synthesis for Vision Language Model Training**
Xiaomin Yu, Pengxiang Ding, Wenjie Zhang, **Siteng Huang**, Songyang Gao, Chengwei Qin, Kejian Wu, Zhaoxin Fan, Ziyue Qiao, Donglin Wang

[P6] **Exploring the Evolution of Physics Cognition in Video Generation: A Survey**
Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, **Siteng Huang**✉, and Donglin Wang✉

[P7] **Humanoid-VLA: Towards Universal Humanoid Control with Visual Integration**
Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, Yuefan Wang, Huaicheng Zhou, Wenshuo Feng, Jiacheng Liu, **Siteng Huang**, Donglin Wang

[P8] **Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models**
Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Bo Zheng, Linfeng Zhang, **Siteng Huang**, Honggang Chen

[P9] **Score and Distribution Matching Policy: Advanced Accelerated Visuomotor Policies via Matched Distillation**
Bofang Jia, Pengxiang Ding, Can Cui, Mingyang Sun, Pengfang Qian, **Siteng Huang**, Zhaoxin Fan, Donglin Wang

[P10] **Filter, Correlate, Compress: Training-Free Token Reduction for MLLM Acceleration**
Yuhang Han*, Xuyang Liu*, Zihan Zhang, Pengxiang Ding, Donglin Wang, Honggang Chen, Qingsen Yan, **Siteng Huang**✉

[P11] **Sparse-Tuning: Adapting Vision Transformers with Efficient Fine-tuning and Inference**
Ting Liu, Xuyang Liu, **Siteng Huang**, Liangtao Shi, Zunnan Xu, Yi Xin, Quanjun Yin, Xiaohong Liu

[P12] **Focus-Consistent Multi-Level Aggregation for Compositional Zero-Shot Learning**
Fengyuan Dai, **Siteng Huang**, Min Zhang, Biao Gong, Donglin Wang

Professional Services

- **Journal reviewer** for RA-L, TNNLS, TCSVT, ACM TIST, JVCI, CPE
- **Conference reviewer** for CVPR, ICCV, ECCV, AAAI, IJCAI, ACMMM, ICME, ICMR, ACCV, ICPR

Honors & Awards

- Outstanding Graduates, Zhejiang University 2024
- **National Scholarship (Top 1%, highest scholarship from Ministry of Education of China)** 2020
- Postgraduate Academic Fellowship, Zhejiang University 2019–2024
- Excellent Student Scholarship, Wuhan University (Top 5%) 2016–2019