Zangwei Zheng

zhengzangwei@gmail.com · zangwei.dev

EDUCATION

National University of Singapore Aug. 2021 – May 2025
Ph.D. in Computer Science. Advisor: Prof. Yang You *Singapore*
- Research Achievement Award of NUS

Nanjing University Sep. 2017 – Jun. 2021
B.S. in Computer Science and Technology, National Elite Program in Computer Science *Jiangsu, China*
- **GPA:** 4.61/5.00 (92.2/100, top 2%)

ACADEMIC RESEARCH EXPERIENCE

HPC-AI Tech Mar. 2024 – Mar. 2025
Team lead & first author of the video generation model **Open-Sora 29k stars** *Singapore*
- Led development and distributed training of a large-scale video generation model from scratch.
- Designed data pipeline and training framework supporting rectified flow, temporal VAE, dynamic resolution, and image-conditioned generation.
- Built scalable training and inference stack for large-scale video generation.

ByteDance Jun. 2021 – Jun. 2022
Research intern, in charge of large batch training for click-through rate prediction model *Singapore*
- Redesigned the training pipeline for large-scale CTR prediction models by converting asynchronous training to synchronous distributed training.
- Developed the CowClip algorithm enabling stable training with batch sizes up to 512k, significantly accelerating training efficiency and improving model AUC (AAAI 2023 Distinguished Paper Award).

National University of Singapore (HPC-AI Lab) Aug. 2021 – May 2025
Ph.D. student, supervised by Prof. Yang You *Singapore*
- Designed an LLM inference pipeline that predicts response length to improve scheduling efficiency, reducing tail latency and improving throughput in large-scale inference systems (NeurIPS 2023).
- Developed continual learning techniques for vision-language models to mitigate zero-shot degradation, improving cross-domain generalization in multi-task learning settings (ICCV 2023).

University of California, Berkeley (iCyPhy, DOP Center) Apr. 2020 – May 2021
Research intern, supervised by Prof. Alberto Sangiovanni-Vincentelli & Dr. Xiangyu Yue *(remote) CA, US*
- Developed self-supervised clustering methods for few-shot domain adaptation, improving cross-domain generalization in visual recognition tasks (CVPR 2021).
- Designed scene-aware learning techniques for radar object detection, including improved backbone architectures and data augmentation strategies.

SELECTED PUBLICATIONS

1. **Open-Sora 2.0: Training a Commercial-Level Video Generation Model in $200k**
Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, et al. **arXiv 2025**
2. **Open-Sora: Democratizing Efficient Video Production for All** Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, Yang You **arXiv, 2024**
3. **Response Length Perception and Sequence Scheduling: An LLM-Empowered LLM Inference Pipeline**
Zangwei Zheng, Xiaozhe Ren, Fuzhao Xue, Yang Luo, Xin Jiang, Yang You **NeurIPS 2023**
4. **Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models**
Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, Yang You **ICCV 2023**
5. **CAME: Confidence-guided Adaptive Memory Efficient Optimization** Yang Luo, Xiaozhe Ren, Zangwei Zheng, Xin Jiang, Zhuo Jiang, Yang You **Distinguished Paper Award (0.8%), ACL 2023**
6. **CowClip: Reducing CTR Prediction Model Training Time from 12 hours to 10 minutes on 1 GPU** Zangwei Zheng, Pengtai Xu, Xuan Zou, Da Tang, Zhen Li, Chenguang Xi,

Peng Wu, Leqi Zou, Yijie Zhu, Ming Chen, Xiangzhuo Ding, Fuzhao Xue, Ziheng Qing, Youlong Cheng, Yang You **Distinguished Paper Award (0.1%), AAAI 2023**

7. **Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation** Xiangyu Yue*, Zangwei Zheng*, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, Alberto Sangiovanni-Vincentelli **CVPR 2021**

INDUSTRY EXPERIENCE

HPC-AI Tech Mar. 2025 – Present

Team lead of the AI video generation platform **Video-Ocean** *Singapore*

- Led development of Video-Ocean, a SaaS platform for AI video generation integrating multiple third-party generative models.
- Built and deployed a production AI application end-to-end, covering system design, model integration, and product development.
- Designed the system architecture for integrating multiple generative model APIs and orchestrating multimodal generation pipelines.
- Developed video generation agents that automate multi-step content generation for minute-level videos.

SKILLS

Languages	Python, JavaScript (TypeScript), C/C++, Go, Rust, SQL, LaTeX
Deep Learning	PyTorch, TensorFlow, Deepspeed, SGLang
Machine Learning	OpenCV, Scikit-learn, NumPy, FFmpeg
Frontend	React, Next.js
Backend	FastAPI, go-zero
DevOps	Docker, CI/CD (Gitlab & Github), Git, Linux