

Zangwei Zheng

✉ zhengzangwei@gmail.com · 🏠 zangwei.dev

EDUCATION

National University of Singapore

Ph.D. in Computer Science, supervised by Prof. Yang You

Aug. 2021 – May. 2025

Singapore

- Research Achievement Award of NUS

Nanjing University

B.S. in Computer Science and Technology, National Elite Program in Computer Science

Sep. 2017 – Jun. 2021

Jiangsu, China

- **GPA:** 4.61/5.00 (92.2/100, top 2%)

RESEARCH INTEREST

Video Generation: pretraining, alignment and agent.

Efficient Machine Learning: deep learning optimizer and efficient inference.

INDUSTRY EXPERIENCE

HPC-AI Tech

Mar. 2025 – Present

Singapore

Team lead of the AI video generation platform Video-Ocean

- Coordinate cross-functional collaboration among design, development, and operations teams.
- Oversee product quality assurance and ensure smooth release processes.

◦ Lead algorithm development for the **Video Agent**, capable of generating minute-level videos.

HPC-AI Tech

Mar. 2024 – Mar. 2025

Singapore

Team lead & first author of the video generation model Open-Sora  27k stars

- Design, develop and train Transformer-based video generation model from scratch.

◦ Design and develop the data processing pipeline. Incorporate features including rectified flow, temporal VAE, image-conditioned generation, dynamic resolution support, etc.

ByteDance

Jun. 2021 – Jun. 2022

Singapore

Research intern, in charge of large batch training for click-through rate prediction model

- Transformed the asynchronous CTR training model into the large-scale synchronous training framework.
- Deployed CowClip algorithm with batch size 512k and improved the AUC of CTR prediction (**AAAI 2023**).

ACADEMIC RESEARCH EXPERIENCE

National University of Singapore (HPC-AI Lab)

May 2019 – Present

Singapore

Ph.D. student, supervised by Prof. Yang You

- Large language model inference acceleration by predicting response length and sequence scheduling.
- Continual learning of vision-language model to prevent zero-shot performance degradation.

University of California, Berkeley (iCyPhy, DOP Center)

Apr. 2020 – May 2021

(remote) CA, US

Research intern, supervised by Prof. Alberto Sangiovanni-Vincentelli & Dr. Xiangyu Yue

- Few-shot Domain Adaptation via Self-supervised Learning with Clustering
- Proposed scene-aware learning with better backbones and data augmentations for radar object detection.

SELECTED PUBLICATIONS

1. **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
2. **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
3. **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
4. **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

5. **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
6. **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

SKILLS

Languages	Python, JavaScript (TypeScript), C/C++, Go, L ^A T _E X
Machine Learning	PyTorch, TensorFlow, ColossalAI, vLLM, OpenCV, Scikit-learn, NumPy, FFmpeg
Frontend	React, Next.js, Tailwind CSS, Vercel, Figma, SEO, SEM
Backend	FastAPI, Postgres
DevOps	Docker, Terraform, CI/CD (Gitlab & Github), Grafana, Agile Development