# Zangwei Zheng

zhengzangw@gmail.com · zhengzangw.github.io

## EDUCATION

**National University of Singapore** Aug. 2021 – Jun. 2025 (expected)
*Ph.D. in Computer Science, supervised by 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%)

## INDUSTRY EXPERIENCE

**HPC-AI Tech** Mar. 2024 – Present
*Research intern, team lead & first author of the video generation model* **Open-Sora 22k stars** *Singapore*
- Development of Video Generation Models: Led the design and scaling of video generation models, implementing advanced attention mechanisms, diffusion loss functions, and a unified framework for image-to-video generation, video continuation, and character identity consistency.
- Training Strategies, Evaluation, and Monitoring: Designed phased training strategies to improve stability, built automated and human evaluation, and developed a system for real-time monitoring and automatic recovery.
- Data Processing: Collected and processed large-scale video datasets with multi-threaded techniques, including segmentation, scoring, and annotation, ensuring high-quality data for training.
- Training Optimization: Enhanced single-machine training performance, improved multi-machine communication efficiency, and resolved memory leakage issues in video loading libraries.
- Distributed Training Setup: Optimized bare-metal systems for distributed training, improving storage, file systems, and NCCL parameters for efficient large-scale operations.
- Commercial Deployment: Successfully launched the commercial version of the model, VideoOcean, delivering scalable video generation capabilities for real-world applications.

**ByteDance** Jun. 2021 – Jun. 2022
*Research intern, in charge of large batch training for click-through rate prediction model* *Singapore*
- 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
*Ph.D. student, supervised by Prof. Yang You* *Singapore*
- Large language model inference acceleration by predicting response length and sequence scheduling.
- Continual learning of vision-language model to prevent zero-shot performance degradation.
- Acceleration of recommendation system training by large batch training.
- Introduce prompt learning for domain generalization with vision transformer.

**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*
- 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

**Large-scale Deep Learning Optimization**

1. **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**
2. **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**

3. **Helen: Optimizing CTR Prediction Models with Frequency-wise Hessian Eigenvalue Regularization**
Zirui Zhu, Yong Liu, Zangwei Zheng, Huifeng Guo, Yang You **WWW 2024**

**Efficient Machine Learning**

1. **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**
2. **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**
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. **InfoBatch: Lossless Training Speed Up by Unbiased Dynamic Data Pruning** Ziheng Qin, Kai Wang,
Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Daquan Zhou, Yang You **ICLR 2024**

**Large Language Model Scaling**

1. **Openmoe: An early effort on open mixture-of-experts language models** Fuzhao Xue, Zian Zheng, Yao
Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, Yang You **ICML 2024**
2. **To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis**
Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, Yang You **Neurips 2023**
3. **A Study on Transformer Configuration and Training Objective**
Fuzhao Xue, Jianghai Chen, Aixin Sun, Xiaozhe Ren, Zangwei Zheng, Xiaoxin He, Yongming Chen,
Xin Jiang, Yang You **ICML 2023**

## SKILLS

| | |
|---|---|
| **Languages** | Python, C, C++, LaTeX |
| **Deep Learning** | PyTorch, TensorFlow, Huggingface, ColossalAI, Deepspeed |
| **Tools** | Pandas, NumPy, Scikit-learn, Gradio, OpenCV, FFmpeg, SQL, Git, Linux |