Kai Yi

williamyi96@gmail.com ◇ Google Scholar ◇ kaiyi.me ◇ (+86) 13772103675
1180 Discovery Wy, Sunnyvale, CA 94089, United States

SUMMARY

I am a Research Scientist at Meta Sunnyvale, working on model compression and inference acceleration. I received my Ph.D. in Computer Science from KAUST, supervised by Prof. Peter Richtárik, and I am expected to graduate in 2025. I have interned at Sony AI, Vector Institute, Tencent AI Lab, and SenseTime Research. My research primarily focuses on **Centralized/Federated LLM Compression**. As the primary author, I have co-authored over 20 papers, accumulating 550+ citations. My work is highly interconnected, featuring significant projects such as the LLM post-training compression algorithms **SymWanda**, **GaLoreEF**, and **PV-Tuning** (*NeurIPS Oral*), with more on the way; communication-efficient federated learning methods **CohortSqueeze** (*NeurIPS-W Oral*), **FedP3** (*ICLR*), and **EF-BV** (*NeurIPS*); and multimodal language model projects **DACZSL** (*ICCV-W*), **HGR-Net** (*ECCV*), and **VisualGPT** (*CVPR*).

EXPERIENCE

Research Scientist @ Meta. Sunnyvale, CA Aug 2025 - Now

- Working on neural compression and model inference acceleration.

Research Intern @ Sony AI. Tokyo, Japan Jun 2023 - Sep 2023

- Innovated federated learning strategies for one-for-all foundation models, leading to significant advancements detailed in FedP3 (ICLR'24).

Research Intern @Vector Institute. Remote May 2023 - Sep 2023

- Federated stochastic bilevel optimization and Newton methods for bilevel optimization. Designed efficient fully single-loop variance reduced methods based on L-SVRG for stochastic bilevel optimization.

Research Intern @ Tencent AI Lab. Shenzhen, China Dec 2020 - Apr 2021

- Developed ML algorithms tailored for bioinformatics data, enhancing commercial products at Tencent.

Research Intern @ Sensetime Group Limited. Beijing, China Mar 2019 - Jun 2019

- Created accurate and fast object detection methods for commercial embedded chips at SenseTime.

Research and Engineering Intern @ IAIR-XJTU. Xi'an, China Jul 2017 - Feb 2019

- Developed cognition-based small object detection methods for autonomous driving and enhancing the Pioneer I autonomous vehicle.

EDUCATION

King Abdullah University of Science and Technology (KAUST) Dec 2021 - Jun 2025
Ph.D. Candidate supervised by Prof. Peter Richtárik
Research Interests: LLM Compression, Federated Learning, Distributed Optimization
Dissertation: Strategies for Improving Communication Efficiency in Distributed and Federated Learning: Compression, Local Training, and Personalization

King Abdullah University of Science and Technology (KAUST) Sep 2020 - Dec 2021
M.S. of Vision-CAIR, supervised by Prof. Mohamed Elhoseiny
Research Interests: Zero-Shot Learning, Vision and Language
Thesis: Domain-Aware Continual Zero-Shot learning

Xi'an Jiaotong University (XJTU), Xi'an, China Aug 2015 - Jun 2019
B.S. of Software Engineering
Thesis: Accurate Object Detection and Weakly-Supervised Perception in Complex Scenes, supervised by Prof. Nanning Zheng and rated as A+ (Top 1%)

HIGHLIGHTED PUBLICATIONS

[1] **Kai Yi**, Peter Richtárik. Symmetric Post-Training Pruning and Training-Free Fine-Tuning for Large Language Models. **ICLRW**, 2025.

[2] **Kai Yi**, Timur Kharisov, Igor Sokolov, Peter Richtárik. Cohort Squeeze: Beyond a Single Communication Round per Cohort in Cross-Device Federated Learning. **NeurIPSW (Oral)**, 2024.

[3] Vladimir Malinovskii, Denis Mazur, Ivan Ilin, Denis Kuznedelev, Konstantin Pavlovich Burlachenko, **Kai Yi**, Dan Alistarh, Peter Richtárik. PV-Tuning: Beyond Straight-Through Estimation for Extreme LLM Compression. **NeurIPS (Oral)**, 2024.

[4] **Kai Yi**, Nidham Gazagnadou, Peter Richtárik, Lingjuan Lv. FedP3: Federated Personalized and Privacy-friendly Network Pruning under Model Heterogeneity. **ICLR**, 2024.

[5] **Kai Yi**, Xiaoqian Shen, Yunhao Gou, Mohamed Elhoseiny. Exploring Hierarchical Graph Representation for Large-Scale Zero-/Few-Shot Image Classification. **ECCV**, 2022.

[6] Laurent Condat, **Kai Yi**, Peter Richtárik. A Unified Theory of Error Feedback and Variance Reduction Mechanisms for Controlling Biased and Unbiased Gradient Compressors in Distributed Optimization. **NeurIPS**, 2022.

[7] Jun Chen, Han Hao, **Kai Yi**, Boyang Li, Mohamed Elhoseiny. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. **CVPR**, 2022.

TEACHING & SERVICES

Reviewer:

T-PAMI, T-MC, IJCV, CVIU, T-IP, T-SP, T-NNLS
NeurIPS'22-25, ICLR'23-25, ICML'22-25, AISTATS'23,25, CVPR'22-25, ICCV'23
ECCV'22,24, AAAI'22-24, WACV'21-25, BMVC'20-23, ITSC'20-21, IV'18-21

Teaching Assistant:

CS283: Deep Generative Modeling (KAUST)
Introduction to Machine Learning, Computer Architecture (XJTU)

TALKS

- Oral presentation: Cohort-Squeeze, NeurIPS'24 FL@FM workshop.	2024.12.15
- Oral presentation on PV-Tuning, representing the group at NeurIPS'24.	2024.12.12
- Invited talk at SonyAI presenting our federated pruning project.	2023.09.29
- Invited talk at SonyAI-PPML talking about Accelerated LT Methods in FL.	2023.08.23
- Invited talk at Vector Institute Demo Day talking "Optimal and Efficient Variance Reduced Methods for Stochastic Bilevel Optimization"	2023.08.17
- Invited presenter at KAUST VCC Open House 2023 talking ProxSkip-VR.	2023.03.02
- Spotlight talk of EF-BV at KAUST Rising Stars in AI Symposium 2023.	2023.02.21
- Representing our group to present ProxSkip-VR at KAUST VCC Showcase Event.	2023.01.29
- Invited speaker at ECCV2022-AI TIME talking about our HGR-Net.	2022.12.07
- Spotlight talk of CIZSL++ at KAUST Conference on Artificial Intelligence.	2021.04.28

AWARDS & HONORS

- KAUST Graduate Scholarship	2020-
- Outstanding Graduates of XJTU (top 5%)	2019
- Zeng Xianzi Scholarship (37/4100, top 0.9%)	2016-2018
- Candidate of 6th Excellent Model Student of XJTU (3/37)	2018
- Excellent Student Award (top 5%) of XJTU	2016-2018