Yiyang CAI

(+86)151-2007-1696, (+852)6150-4684 | ycaicj@connect.ust.hk
YIYANGCAI | Google Scholar
Hong Kong University of Science and Technology - Clear Water Bay - Hong Kong

Education

- **Hong Kong University of Science and Technology** *Sep. 2024 - Current*
  Ph.D. Student of Independent Interdisciplinary Program (Supervisor: Prof. Wenhan LUO and Prof. Yike GUO) Kowloon, HKSAR
  ◦ Research Interests: Computer Vision, Generative Models, Personalized Generation

- **University of California, Berkeley** *Aug. 2020 - Dec. 2021*
  Master of Engineering, Visual Computing and Computer Graphics, (Supervisor: Prof. Brian BARSKY) Berkeley, CA, United States
  ◦ GPA: 3.65/4.00

- **Beihang University** *Aug. 2016 - Jun. 2020*
  Bachelor of Engineering, Automation (Shenyuan Honor School) Beijing, China
  ◦ GPA: 3.81/4.00 (Outstanding Graduate Honor)

Publications

[1] **Y. Cai**, Z. Jiang, Y. Liu, C. Jiang, W. Xue, Y. Guo & W. Luo . Foundation Cures Personalization: Improving Personalized Models' Prompt Consistency via Hidden Foundation Knowledge. The 39th Annual Conference on Neural Information Processing Systems, 2025.

[2] C. Jiang, C. Chan, **Y. Cai**, Y. Liu, W. Xue, & Y. Guo (2025). Graceful Forgetting in Generative Language Models. Empirical Methods in Natural Language Processing (Main Conference), 2024.

[3] W. Cheng, W. Zhang, H. Shen, **Y. Cai**, X. He, K. Lv, & Y. Liu . Optimize weight rounding via signed gradient descent for the quantization of LLMs. Empirical Methods in Natural Language Processing, 2024.

[4] H. Chang, H. Shen, **Y. Cai**, X. Ye, Z. Xu, W. Cheng, K. Lv, W. Zhang, Y. Lu & H. Guo. Effective quantization for diffusion models on CPUs. NeurIPS 2023 Workshop on Diffusion Models, 2023.

[5] W. Cheng, **Y. Cai**, K. Lv, & H. Shen. TEQ: Trainable equivalent transformation for quantization of LLMs. arXiv preprint arXiv:2310.10944, 2023.

[6] L. Tang*, **Y. Cai***, J. Liu, Z. Hong, M. Gong, M. Fan, J. Han, J. Liu, E. Ding, & J. Wang. Few-shot font generation by learning fine-grained local styles. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7895-7904). ("*" means equal contribution), 2022.

Projects

- **Training-free Prompt Following Optimization for Facial Personalization Models** *Sep. 2024 - Apr. 2025*
  Keyword: Personalized Diffusion Models, Prompt Following, Training-free [GitHub]
  ◦ Completed a comprehensive literature review of facial personalization approaches.
  ◦ Identified the current limitation shared by state-of-the-art facial personalization methodologies: poor prompt following performance on fine-grained facial attributes (e.g., hair style, accessories, expression).
  ◦ Proposed a dual-denoising branch in personalization models (w./w.o. identity information) with an optimized self-attention modules that transfers satisfying facial attributes into personalization denoising process.
  ◦ Validated the proposed method on popular personalization models built upon different foundation models (e.g., SDv1.5, SDXL, FLUX.1-dev) and finished a paper submission.

Working Experience

- **Intel Inc. (China)** [🌐] *Jun. 2022- Jun. 2024*
  Machine Learning Engineer, Group of Data Center and Artificial Intelligence Shanghai, China
  ◦ Developed and maintained key features of **Intel Neural Compressor**, including quantization and pruning for diffusion models and large language models.
  ◦ Developed efficient GPTQ API for a large variety of LLMs and VLMs (e.g., MPTs, Qwens, ChatGLMs, Baichuan, LLaVA). Designed pruning API to support different orthogonal configurations (structures, criterion, schedulers, etc.). Validated API's function on a wide range of model, including CV, NLP models and popular LLMs.

- **Baidu Inc.** [🌐] *Sep. 2020 - Nov. 2021*
  Algorithm Intern, VIS Group Beijing, China
  ◦ GAN-based style transfer research for few-shot Chinese font generation: given a small set of style-specified glyphs ($\leq$ 100 examples), generated the entire personalized font ($\approx$ 20K instances). Applied a novel attention machanism to improve the model's ability of capturing fine-grained style and spatial feature representations from glyphs. Designed character decomposition strategy to boost efficiency of learning characters' component representations.

SKILLS

- **Programming Languages:** Python, C++
- **Deep Learning & Computer Vision Tools:** PyTorch, DeepSpeed, Matplotlib, OpenCV and Pandas
- **Language Proficiency:** TOEFL: 104 (Reading 28 + Listening 25 + Speaking 26 + Writing 25)

TEACHING EXPERIENCE

- **Teaching Assistant** *Feb. 2025 - May. 2025*
  TA duty for EMIA-4110 (Practical Machine Learning) (Teacher: Prof. Wenhan LUO) HKUST
  - Delivered several presentations of using Colab and PyTorch to undergraduate students.
  - Supported the professor's course preparation of image generation tutorials.