Yiran Huang

Munich, Germany | yiranh97@gmail.com | https://yiranhuangirene.github.io/ | +49 152 5915 6110 / +86 15995326157

Research Profile

PhD researcher at the Technical University of Munich, advised by Prof. Zeynep Akata, working on **multimodal LLMs**, with a focus on **mechanistic interpretability** and **efficient post-training**. First- and co-author publications at ICML (Spotlight), ICLR, IJCV, GCPR (Oral), and NeurIPS. Experienced in leading projects end-to-end, from research question to large-scale training and publication. Seeking a **Summer/Fall 2026 research internship** in LLM/MLLM research.

Education

Technical University of Munich — PhD in Computer Science — Aug 2023 – Present
Chair of Interpretable and Reliable Machine Learning; advised by Prof. Zeynep Akata. Member of IMPRS-IS and MCML.

Technical University of Munich — M.Sc. Robotics, Cognition, Intelligence — Oct 2020 – Jul 2023

Tongji University — B.Sc. Vehicle Engineering — Sep 2015 – Jul 2019

Research Experience

Technical University of Munich — PhD Researcher, Chair of Prof. Zeynep Akata — Aug 2023 – Present

- Lead first-author research on multimodal LLMs spanning mechanistic interpretability, in-context learning, and efficient post-training; published at ICML (spotlight), ICLR, IJCV, and GCPR (oral).
- Own projects end-to-end: problem formulation, experimental design, large-scale training/eval, paper writing, and rebuttal — across model families including LLaVA, Bunny, InternVL, and Qwen2.5-VL.
- Built reusable training and evaluation infrastructure on multi-node GPU clusters using PyTorch, Hugging Face, PEFT, DeepSpeed, FlashAttention, and SLURM.
- Collaborate with international co-authors (Tübingen, Trento, UC Berkeley) and co-supervise MSc thesis and seminar students.

Selected Publications

Dissecting Multimodal In-Context Learning: Modality Asymmetries and Circuit Dynamics in Modern Transformers
Y. Huang, K. Roth, Q. Bouniot, W. Xu, Z. Akata. *ICML 2026, Spotlight*.

- First mechanistic account of multimodal in-context learning in modern transformers; establishes a controlled testbed for analyzing how architectural choices and data statistics shape ICL across modalities.
- Reveals a learning asymmetry between modalities, characterizes the underlying circuit dynamics, and validates the findings on production-scale MLLMs.

Structural Pruning of Large Vision-Language Models: Pruning Dynamics, Recovery, and Data Efficiency
Y. Huang, L. Thede, M. Mancini, W. Xu, Z. Akata. *IJCV 2026*.

- Studies layerwise and widthwise structural pruning in open-source vision-language models.
- Shows that SFT + hidden-state distillation can retain more than 95% of original performance using only 5% of recovery data.

Towards Understanding Multimodal In-Context Learning Through Classification Tasks
Y. Huang, K. Roth, Q. Bouniot, W. Xu, Z. Akata. *WCTD @ NeurIPS 2025*.

Investigating Structural Pruning and Recovery Techniques for Compressing MLLMs
Y. Huang, L. Thede, M. Mancini, W. Xu, Z. Akata. *GCPR 2025, Oral*.

Revealing and Reducing Gender Biases in Vision and Language Assistants (VLAs)
L. Girrbach, S. Alaniz, **Y. Huang**, T. Darrell, Z. Akata. *ICLR 2025*.

- Evaluates gender bias in 22 MLLMs across skills and occupations; finds that fine-tuning-based debiasing methods achieve the best trade-off between debiasing and retaining performance.

Additional Experience

Technical University of Munich — Seminar Instructor — Apr 2024 – Present
Designed and led a graduate seminar on **MLLMs and foundation models**; mentored students in critical reading and evaluation of research papers from both reviewer and practitioner perspectives.

Bosch Sensortec GmbH — Machine Learning Intern — Apr 2022 – Sep 2022
Developed a lightweight cascaded CNN for real-time human activity recognition; applied pruning and quantization and deployed the model in a C++ embedded framework on ultra-low-power microcontrollers.

Skills

Research Areas: Multimodal LLMs, in-context learning, mechanistic interpretability, post-training, pruning, distillation, bias evaluation.
ML Systems: PyTorch, JAX, Hugging Face, PEFT, Accelerate, DeepSpeed, FlashAttention, vLLM, SLURM, Docker, Weights & Biases.
Programming: Python, C++.
Languages: Chinese (native), English (fluent), German (intermediate).