Zicong Tang

Tel: +86 187 727 22454 **Email:** tangzc@whu.edu.cn

Education

Wuhan University (WHU) **Wuhan, China**
School of Computer Science B.S. (Ranked 16th by US News) Sepetember 2022–present
Score&GPA: 94.15/100, Rank**1**/204; **3.97**/4.00, Rank**2**/204
Core courses: Advanced Mathematics 100, Linear Algebra 100, Discrete Mathematics 98, Operating Systems 98, Software Design and Architecture 98, Principles of Compiler 98, Advanced Programming Language 97, Database Systems 96
Research Area: Large Language Models, Multi-Modal Large Language Models, KV Cache Compression, Multi-Modal Representation, Multi Token Prediction

Pubilications

CoViPAL: Layer-wise Contextualized Visual Token Pruning for Large Vision-Language Models **Reprint**
- **Zicong Tang**, ZiyangMa, SuqingWang, Zuchao Li, Lefei Xhang, Hai Zhao, Yun Li, Qianren Wang
- Underreview, ACL ARR 2025 May, **First author** .

SpindleKV: A Novel KV Cache Reduction Method Balancing Both Shallow and Deep Layers
- **Zicong Tang**, Luohe Shi, Zuchao Li, Baoyuan Qi, Guoming Liu, Lefei Zhang
- Main conference of **ACL2025, first author**.

Research Experience

Layer-wise Contextualized Visual Token Pruning for Large Vision-Language Models Feb. 2025 - present
Wuhan University, Sigma Lab Advisor: Zuchao Li
- We are the first to identify inherent redundancy in visual tokens, which is layer-irrelevant.
- We train a compact classifier on a small, negligible dataset and effectively reduce this redundancy.
- Our method reduces decoding time by 60% and prunes 75% of visual tokens, resulting in only a slight performance degradation.

LLM KV Cache Reduction Method Balancing Both Shallow and Deep Layers Sept. 2024 - Dec. 2024
Wuhan University, Sigma Lab Advisor: Zuchao Li
- Identify two forms of redundancy in the KV Cache: attention sparsity and KV constitutional similarity, with the former being more pronounced in deep layers and the latter in shallow layers.
- Propose a pioneering method that combines eviction and merging: employ an attention-weight-based eviction method and a codebook-based replacement method to mitigate the layer-wise redundancy and achieve SOTA performance.
- Address a commonly overlooked challenge: our method tackles the issue faced by eviction-based methods when integrating with GQA and achieves surprising performance on GQA models.

Honors and Awards

- **China National Scholarship**, Top 2% nationwide Oct. 2023
- First-class Scholarship, Twice, Top 5% schoolwide Oct. 2023 & Oct. 2024
- Merit Student, Twice, Top 5% schoolwide Oct. 2023 & Oct. 2024
- Silver Medal, Team Leader, Regional, Chinese Collegiate Computing Competition (4C) Jun. 2024

Projects

Computer Vision & Nature Language Processing (Course Project) June. 2024
- Perform image classification on the CIFAR-10 dataset using the classic LeNet model and further improve the model structure to enhance performance.
- Fine-tune the chatglm2-6b model on the AdvertiseGen dataset with QLora, 4-bit quantization of the base model.

CPU Design for RISC-V Instruction Set (Course Project) Mar. 2024
- Used the Verilog language to design and implement a five-stage pipeline CPU, including IF/ID/EX/MEM/WB stages.
- Implemented the decoding and execution of the RISC-V instruction set, including arithmetic, load/store, branch, etc.

Service & Activities

- Outstanding Youth Volunteer with 132 hours of service, Top 5% of the Computer Department — Aug. 2024
- Minister of the Secretariat Department of the Taekwondo Association at Wuhan University — Oct. 2023 - present
- Outstanding Teaching Assistant, serve as a teaching assistant for Data Structures. — Mar. 2024 - Jun. 2024
- Volunteer teaching Physics and Introduction to AI in a high school in Guizhou Province — Jul. 2024 - Aug. 2024