Junhan Zhu

zhujunhan@westlake.edu.cn | alrightlone.github.io | LinkedIn | GitHub

Research Interests

Efficient AI · Computer Vision · Model Compression · Inference Acceleration

Education

Westlake University *Hangzhou, China*
Bachelor of Engineering in Electronic and Information Engineering **Sept. 2023 – Present**

- **GPA: 4.04/4.3 (Rank: 2/23)**
- **Early Admission:** Admitted by skipping the last year of high school due to academic excellence.

Nanyang Technological University (NTU) *Singapore*
Exchange Student in School of Electrical and Electronic Engineering (EEE) **Jan. 2026 – May 2026 (Expected)**

Experience

MMLab, Nanyang Technological University (NTU) *Advisor*: *Prof. Ziwei Liu*
Research Intern **Jan. 2026 - Present**

- **Research Focus:** Inference acceleration for Vision-Language Models (VLMs) and unified multimodal models.

ENCODE Lab, Westlake University *Advisor*: *Prof. Huan Wang*
Research Intern **Mar. 2025 - Nov. 2025**

- **Research Focus:** Model compression and acceleration for diffusion models.
- Proposed **OBS-Diff** by surveying OBS literature and designing a **Timestep-Aware Hessian** and **Module Package** strategy to address iterative error accumulation and minimize calibration costs.
- Developed a versatile framework compatible with **U-Net (SDXL)** and **MMDiT (SD3.5)** architectures, supporting unstructured, semi-structured, and structured pruning within a unified codebase.
- Achieved **training-free, one-shot** pruning on large-scale models, maintaining competitive FID/CLIP scores even at **30% structured sparsity** and significantly outperforming SOTA methods. Accepted to **ICLR 2026** as **First Author**.

Publications

* Corresponding author. † Equal contribution (Co-first author).

[1] OBS-Diff: Accurate Pruning For Diffusion Models in One-Shot
Junhan Zhu, Hesong Wang, Mingluo Su, Zefang Wang, Huan Wang* ICLR 2026 *(Lead Author)* | **Jan. 2026**
[arXiv] [Project] [Code]

- Proposed the novel training-free, one-shot pruning framework for diffusion models via Optimal Brain Surgeon (OBS), achieving SOTA performance across diverse architectures and granularities.

[2] Cross-Resolution Diffusion Models via Network Pruning
Jiaxuan Ren†, ***Junhan Zhu***†, Huan Wang* **Under Review** *(Co-first Author)* | **Nov. 2025**

- CR-Diff repurposes network pruning to enhance generalizability by removing "adverse weights" that cause degradation at non-default resolutions.

Selected Projects

SparAlloc: Modular Framework for LLM Sparsity Allocation
Lead Developer | *Source Code* **May 2025**

- **Topic:** Decoupling the "where to prune" (allocation) from "how to prune" (importance metric) in LLM compression.
- **Implementation:** Built a modular Python framework supporting diverse layer-wise sparsity allocation.
- **Outcome:** Enabled rapid benchmarking of sparsity distributions across Llama-7B, significantly reducing the experimental cycle for finding optimal pruning masks.

Awards

- **Hongyi Scholarship** Dec. 2024
  Westlake University | *Awarded for meaningful social activities*

- **Outstanding Bachelor's Student** Oct. 2025
  Westlake University | *Awarded for outstanding academic performance*
- **Outstanding Undergraduate Scholarship** Oct. 2024 & Oct. 2025
  Westlake University | *Awarded for top undergraduate students (¥2000/year)*
- **Innovation Award** Oct. 2024 & Oct. 2025
  Westlake University | *Awarded for innovative research achievements*

Skills

- **Technical Skills:** Python, C/C++, PyTorch, Hugging Face, Linux, Git, LaTeX
- **Languages:** Chinese (Native), English (Fluent, IELTS 7.0)