Junhan Zhu

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

Brief Intro

I am an undergraduate student at Westlake University, actively seeking Ph.D. opportunities for Fall 2027. My research interests lie in **Efficient AI** and **Computer Vision**, with a focus on developing novel algorithms for model compression and efficient generative models.

Education

Westlake University, Bachelor of Engineering in Electronic and Information Engineering — Sept. 2023 – Present

- Major GPA: 4.04/4.3
- **Selected Coursework:** Data Structures and Algorithms (A+), Calculus (A+), Digital Circuits (A+), Linear Algebra (A), Probability and Statistics (A), Natural Language Processing (A).

Nanyang Technological University (NTU), Singapore — Jan. 2026 – May 2026 (Expected)

Exchange Student in School of Electrical and Electronic Engineering (EEE)

Experience

Visiting Research Student — Mar. 2025 - Present

ENCODE Lab, Westlake University — Advisor: **Prof. Huan Wang**

- Conducting research on efficient generative models, focusing on diffusion model compression via network pruning.

Visiting Research Student — July 2024 - Nov. 2024

TGAI Lab, Westlake University — Advisor: **Prof. Yaochu Jin**

- Proposed a novel Dynamic Time Warping (DTW) based algorithm for optimal threshold selection in aliased signal feature decoding.

Publication

OBS-Diff: Accurate Pruning For Diffusion Models in One-Shot

J. Zhu, H. Wang, M. Su, Z. Wang, H. Wang*

arXiv:2510.06751 | Project Page | GitHub — Oct. 2025

Preprint

- 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.

Cross-Resolution Diffusion Models via Network Pruning

J. Ren†, ***J. Zhu***†, H. Wang*

Under Review at CVPR 2026 (†: Co-first Author) — Nov. 2025

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

Project

SparAlloc: A Modular Framework for Decoupled Sparsity Allocation in LLM Pruning

GitHub — May 2025

- Created a modular framework for benchmarking and flexibly combining diverse sparsity allocation strategies in LLM pruning.

Awards

- **Hongyi Scholarship**, *Westlake University* — Dec. 2024
- **Outstanding Bachelor's Student**, *Westlake University* — Oct. 2025
- **Innovation Award**, *Westlake University* — Oct. 2024 & Oct. 2025

Skills

- **Programming:** Python, PyTorch, C/C++
- **Developer Tools:** Git, LaTeX, Linux Shell
- **Languages:** Chinese (Native), English (Fluent, IELTS 7.0)