

Junhan Zhu

✉ zhujunhan@westlake.edu.cn | 🌐 <https://alrightlone.github.io/> | [in 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).

Experience

Visiting Research Student Mar. 2025 - Present
ENCODE Lab, Westlake University Advisor: **Prof. Huan Wang**
• Proposed *OBS-Diff*, a novel training-free, one-shot pruning framework for diffusion models, supporting diverse architectures and pruning granularities.
• Developed *SparAlloc*, a modular benchmark and toolkit for sparsity allocation algorithms in Large Language Model (LLM) pruning.

Visiting Research Student July 2024 - Nov. 2024
TGAI Lab, Westlake University Advisor: **Prof. Yaochu Jin**
• Investigated foundational principles of Spiking Neural Networks (SNNs).
• Conducted a literature review on the application of AI in chip placement optimization.
• 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 first training-free, one-shot pruning framework for diffusion models, demonstrating broad applicability across diverse architectures and pruning granularities.
• Revitalized the classic Optimal Brain Surgeon (OBS) method for large-scale text-to-image models, achieving state-of-the-art compression performance while maintaining high generative quality, especially at high sparsity regimes.

Project

SparAlloc: A Modular Framework for Decoupled Sparsity Allocation in LLM Pruning May 2025
[GitHub](#)
• Developed a standardized benchmark by collecting and evaluating diverse sparsity allocation algorithms for fair comparison.
• Designed as a modular toolkit to facilitate research by enabling flexible combinations of various pruning algorithms and sparsity allocation methods.

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)
• **Non-Technical:** Communication, Teamwork, Adaptability, Self-Management, Critical Thinking