Towards the Efficient Generative Models

Junhan Zhu
Westlake University
zhujunhan@westlake.edu.cn

Bio: Hi, my name is Junhan ZHU, currently a junior undergraduate from Westlake University, with research interests on Computer Vision and Efficient AI.

- **Efficient AI** (Model Compression and Inference Acceleration)
- **Computer Vision** (VLM, UMM, Omni)

Education

Westlake University

B.Eng in Electronic and Information Engineering

Jul. 2023 - Present (Expected 2027)

Admitted by skipping the final year of high school.

GPA: 4.04 / 4.3 (Rank: 2 / 23)

Achieved A- or above in all major core courses

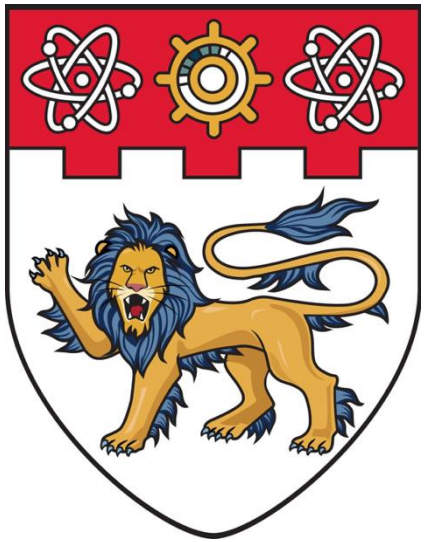
Nanyang Technological University

Exchange Student

Jan. 2026 - Jun. 2026

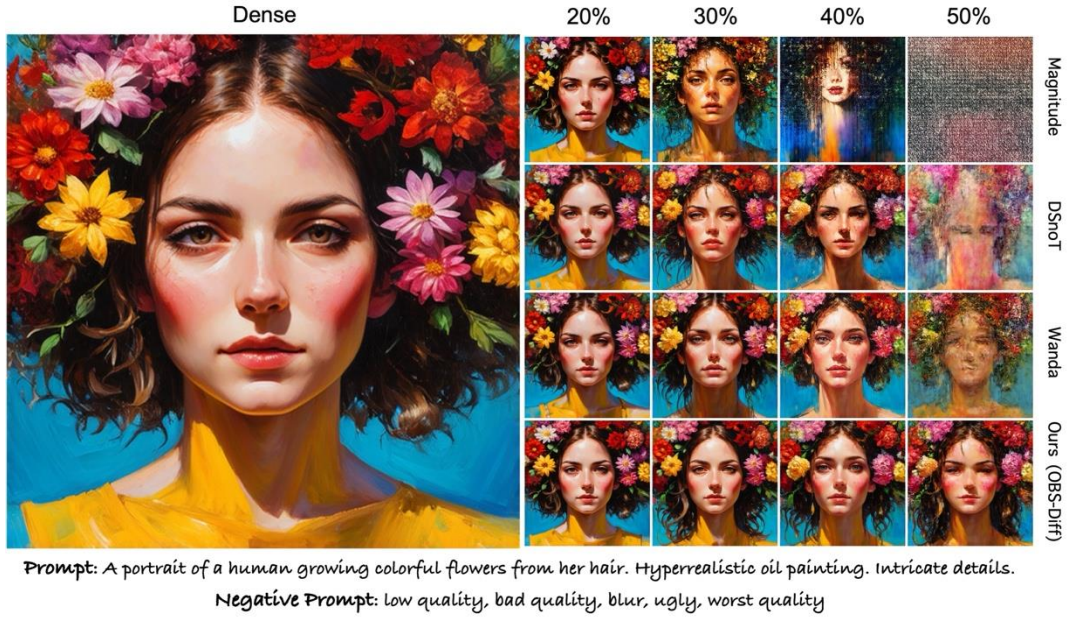
Research Experience

- **Supervisor:** Prof. Huan Wang @ ENCODE Lab, Westlake
- Mar. 2025 - Nov. 2025
- **Research Focus:** Efficient Pruning For Diffusion
- **Outcome:**
  1. SparAlloc (Opensource Project)
  2. OBS-Diff (ICLR 2026, lead author)
  3. CR-Diff (Under Review, co-first author)



- **Supervisor:** Prof. Ziwei Liu @ MMLab, NTU
- Jan. 2026 – Present (Expected June 2026)
- **Research Focus:** Inference Acceleration For VLM

[ICLR 2026] OBS-Diff: Accurate Pruning For Diffusion Models in One-Shot



Qualitative comparison of unstructured pruning methods on the SD3-Medium model.

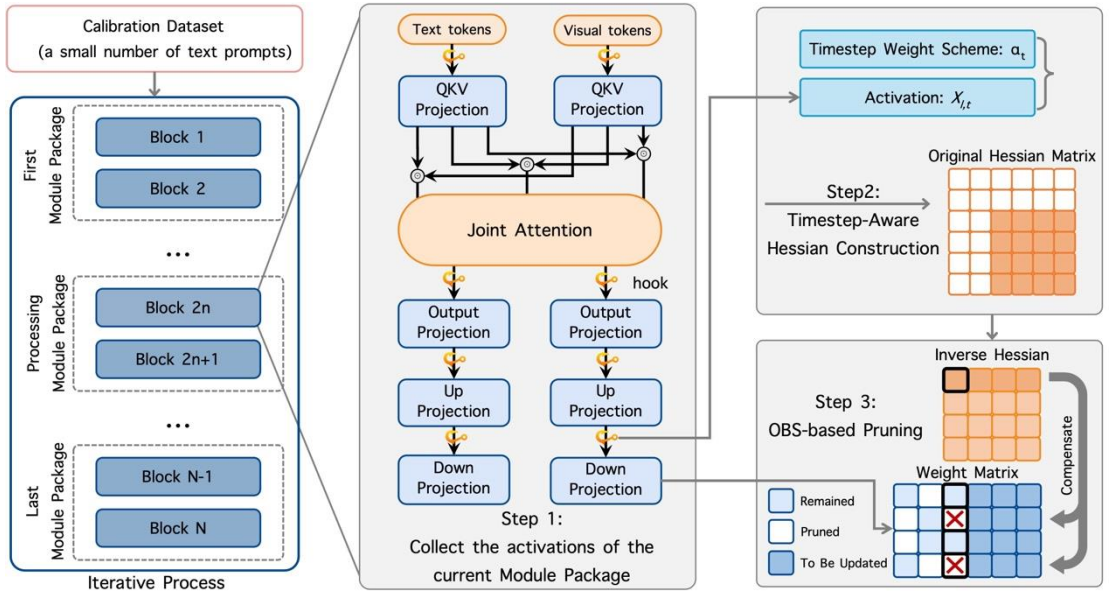


Illustration of the proposed OBS-Diff framework applied to the MMDiT architecture.

- 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.

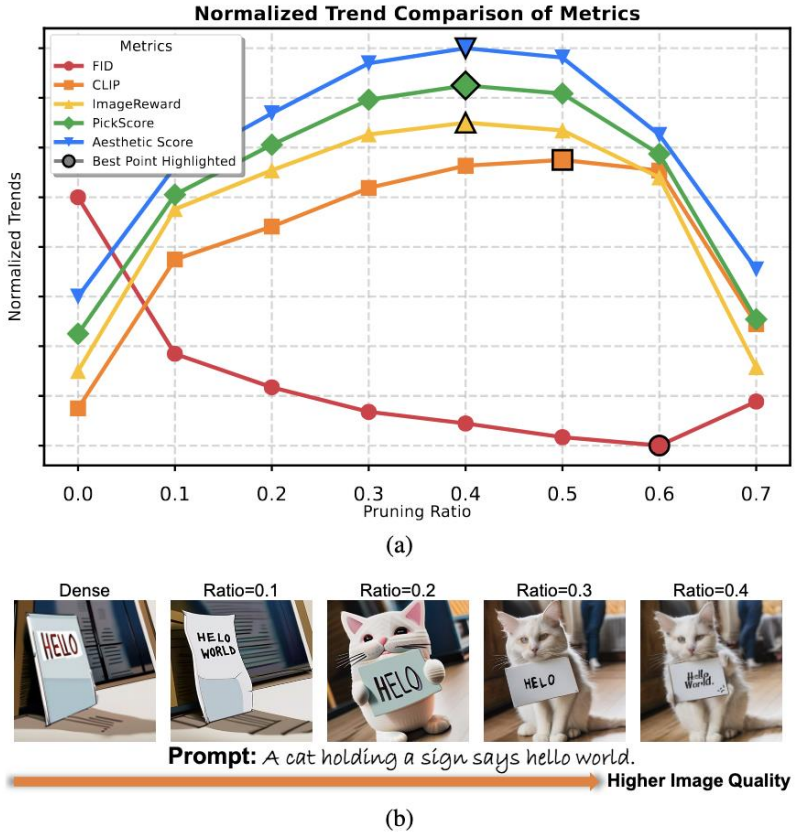
- Developed a versatile framework compatible with **U-Net (SDXL) and MMDiT (SD3.5) architectures**, supporting **unstructured, semi-structured, and structured pruning** within a unified codebase.

- 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.

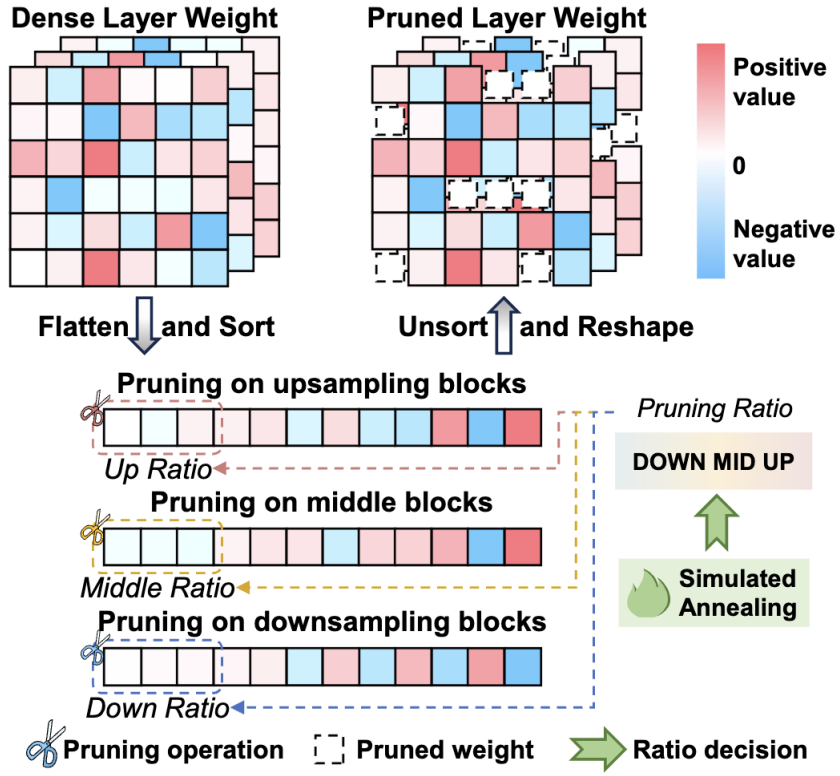
[Under Review] Cross-Resolution Diffusion Models via Network Pruning

- During the development of *OBS-Diff,* identified that **simple magnitude unstructured pruning** could enhance image generation capabilities at non-default resolutions.

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



Effects of magnitude-based unstructured pruning on SDXL at unseen resolution 512×512.



CR-Diff Pipeline

Thank You!