# Zhengyuan Su

su-zy21@mails.tsinghua.edu.cn · timzhengyuansu · https://timzsu.github.io

## Education

**Tsinghua University** 2021 – 2025(expected)

**Yao Class**, Institute for Interdisciplinary Information Sciences (IIIS) *Beijing, China*

*Bachelor of Engineering in Computer Science and Technology* GPA: 3.94/4.00

Award:

**Yao Award** (awarded to Yao Class students with outstanding performances, esp. in academic achievements), 2024; Ubiquant scholarship, 2023; Xiaomi scholarship, 2022

## Publication and Manuscripts

- Zhuo, Y.*, **Su, Z.***, Zhao, C., & Gao, M. (2024). Syno: Structured Synthesis for Neural Operators. Submitted to ASPLOS 2025. `https://arxiv.org/abs/2410.23745`.
- **Su, Z.**, Beyzerov, S., Pang, Q., & Zheng, W. (2024). FABLE: Batched Evaluation on Confidential Lookup Tables in 2PC. Targeting USENIX Security '25. `https://drive.google.com/file/d/1qlz56DrrNXxQ4PSPiFdboFW7ovJgB8r-/view?usp=drive_link`.
- Zhang, J., Fan, G., Wang, G., **Su, Z.**, Ma, K., & Yi, L. (2023, June). Language-assisted 3D feature learning for semantic scene understanding. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 3, pp. 3445-3453). `https://ojs.aaai.org/index.php/AAAI/article/view/25453`.

## Research Experience

**Secure Transformer Inference in Fully Homomorphic Encryption (FHE)** Sep. 2024 – Present

**Advisor:** Prof. Wenting Zheng from CMU *Python, Linux*

- The project aims to accelerate FHE-based secure transformer inference, which can secure users' sensitive data while enabling cloud inference over them using companies' proprietary models.
- Proposed research plan:
  - Estimate benefits to FHE-based secure transformer inference brought by quantization.
  - Construct an automatic framework that assigns appropriate weight precision to balance accuracy and latency.
  - Explore the possibility of generalizing the framework to plaintext weight quantization.

**Leveraging Layer Sensitivity in Outlier-Aware Quantization** Jun. 2024 – Aug. 2024

**Advisor:** Prof. Babak Falsafi from EPFL *Python, Linux*

- The project targets lossless LLM quantization for 4-bit inference, which can substantially reduce the cost of cloud inference service and edge deployment, providing more people with access to LLM.
- Discovered that the impacts of quantizing outliers on the overall perplexity vary across different layers and developed a better quantization strategy.
- Analyzed the mechanism behind sensitivities in-depth, found a close relationship between error propagation and sensitivities, and identified a systematic error pattern existing in sensitive layers that facilitates error propagation.

**Batched Evaluation of Confidential Lookup Tables in 2PC** Sep. 2023 – Jun. 2024

**Advisor:** Prof. Wenting Zheng from Carnegie-Mellon University *C++, Linux*

- The project aims to asymptotically and practically improve the efficiency of confidential LUT evaluation, a key component in privacy-sensitive applications including privacy-preserving ML inference and private data analytics.
- Adapted batching techniques from Private Information Retrieval by designing novel 2PC algorithms and offloading complex procedures to the local side to achieve better performance while ensuring security.
- Achieved asymptotic communication improvement and up to 2 orders of magnitude speedup compared to baselines under various network configurations.
- This project results in a paper targeting USENIX Security '25.

**Structured Synthesis for Neural Operators** Mar. 2023 – Nov. 2023

**Advisor:** Prof. Mingyu Gao from Tsinghua University *Python, C++, Linux*

- The project aims to discover novel neural operators with better accuracy and/or speed. Replacing compute-intensive NN operators with our discovered operators yields higher inference speed with comparable accuracy.
- Proposed the paradigm of neural operator synthesis, an *underexplored* direction orthogonal to existing acceleration techniques including neural architectural search and tensor compilers.
- Built a structural search space for operator synthesis based on view-and-contraction primitives and canonicalization techniques and applied MCTS to search for efficient and expressive operators.
- Achieved $1.10\times$ to $4.73\times$ speedup with less than $2\%$ accuracy loss on ImageNet on multiple vision models.
- This project results in a paper submitted to ASPLOS 2025.

**Language-Assisted 3D Feature Learning** Feb. 2022 – May 2022

**Advisor:** Prof. Li Yi from Tsinghua University *Python, Linux*

- The project aims to enhance 3D semantic scene understanding using free-form natural language descriptions, potentially lowering the data collection cost as descriptions are much easier to collect than 3D scans.
- Parsed objects' attributes and relationships from natural descriptions of 3D scenes as auxiliary training labels.
- Pretrained feature encoders on the labels using pretext tasks, enhancing downstream accuracy by 1-3% across four vision tasks.
- This project results in a paper published in AAAI-23.

## Skills

- Programming Languages: Python, C/C++, Go
- Platform & ML Framework: Linux, PyTorch
- Languages: English - High Working Proficiency (TOEFL 110, Speaking 25), Mandarin - Native speaker