Yixiong Fang

(+86) 131 6210 8292 | kfangyixiong@gmail.com

Education

Shanghai Jiao Tong University (SJTU) Shanghai, China

Bachelor of Engineering in *Software Engineering* Aug. 2021 – Jun. 2025 (Expected)

Major GPA 88.25/100, Major WES GPA 3.86/4.0

Related Courses： Computer Graphics, Computer System Engineering, Computer Vision, Machine Learning, VR/AR & Game Design and Development

Honors and Awards: SJTU President's Award 2023, SJTU Academic Improvement Scholarship 2024, SJTU Merit Undergraduate Scholarship C 2024, Silver Price in Shanghai of the 9th China International College Students' "Internet+" Innovation and Entrepreneurship Competition

Publications

[1] **Yixiong Fang**, Ziran Yang, Zhaorun Chen, et al. "From Uncertainty to Trust: Enhancing Reliability in Vision-Language Models with Uncertainty-Guided Selective Decoding" (submitted to CVPR 2025)

[2] Yalan Lin, Chengcheng Wan, **Yixiong Fang**, et al. "CodeCipher: Learning to Obfuscate Source Code Against LLMs" (submitted to ICLR 2025)

[3] **Yixiong Fang**, Weixi Yang "Efficient Computation of Eigenvalues in Diffusion Maps: A multi-strategy Approach", 2024 7th International Conference on Signal Processing and Machine Learning

Skills

Programming Languages: Python, C++, C#, JAVA, HTML15, Matlab, Axure, JavaScript

Languages: Mandarin (native), English (fluent TOEFL 110 GRE 327)

Research Experience

Optimization of Computing Top-K Eigenvalues and Eigenvectors in Diffusion Mapping **CIS**

Supervised by Prof. David P. Woodruff, Carnegie Mellon University Jul. 2023 - Aug. 2023

- Implemented three optimizations using Gaussian estimation to accelerate the matrix eigenvalue and eigenvector computation and stabilized the speed for calculating Top-K eigenvalues in diffusion maps under 10 seconds, which may cost 2 minutes (some cases) while the accuracy remains .3f
- Performed singular value decomposition on the original matrix; modified a matrix multiplication step in the Arnoldi iteration to reduce computation from O(n^2) to O(ni)
- Constructed eigenvalues and eigenvectors in O(1) complexity using the properties of Markov matrices with k=1, reducing the original computation time to nearly zero

Embedding-Level Code Perturbation for Enhanced Security in Large Language Models (LLMs) **SJTU**

Supervised by Prof. XiaoDong Gu Jul. 2024 - Present

- Design, implement and analysis of the perturbation technique at the embedding level within LLMs, which transformed input tokens into unreadable codes and enhanced security applications; Co-authored paper submitted to The International Conference on Learning Representations 2025
- Conducted extensive testing on LLaMA 7B and LLaMA 13B models, validating the effectiveness of the perturbation method
- Migrated the perturbation method from LLaMA 7B to GPT-3.5, StarCoder, and DeepSeekCoder, demonstrating robust results across different LLM architectures

Mitigating Hallucination in Large Visual Language Models **Stony Brook University**

Supervised by Prof. Jiawei Zhou Apr. 2024 - Present

- Use uncertainty to judge the informative importance of tokens; project visual tokens into language spaces for interpretation; dynamically mask visual tokens for better generation; introduce majority voting strategy in token-level generation
- Developed evaluation metrics, tested on CHAIR, THRONE and MMBench benchmark and achieved satisfactory performance in all metrics across models
- First authored paper submitted to Computer Vision and Pattern Recognition Conference 2025

VR-based High-Realism Recovery System **SJTU**
Supervised by Prof. Xubo Yang Apr. 2024 - Present

- Design and develop a high-realism VR system for recovery using Unity
- Collect data of real lake scene and construct virtual lake scene using **3D-Gaussian**; Adapt to the interaction between physical equipment and virtual water environments, such as pairing a rowing machine with a VR kayaking
- Designed mini-games tailored to patients' needs, such as navigating a kayak toward a marker with the goal of achieving the shortest path, to assist in medical analysis in the impact of **cognitive training** during recovery.

Sign language Translation System Development **SJTU**
Supervised by Prof. Qian Zhang Feb. 2024 - Present

- Implement American Sign Language and Chinese Sign Language translation system, improving the speed and user experience by using ultrasound detection rather than vision methods (ratio hasn't been detected), enabling sequence to sequence translation instead of traditional split to translate methods
- Use FMCW (Frequency Modulated Continuous Wave) method to process ultrasound data, enhancing speed performance than image capturing $ratio, and help users' experience with lighter devices(survey)
- Utilize transformer-based CNN model to process signal data, and Encoder/Decoder architecture to generate translation results

Predict the Prognosis of Patients with Orthodontic-related Alveolar Bone Defects **SJTU**
Supervised by Prof. ZhiGui Ma Sept. 2023 - Present

- Design an efficient and innovative Lasso regression model to predict the prognosis of patients with orthodontic-related alveolar bone defects, which is previously undiscovered
- Achieved unprecedented data collection of CT data from 133 patients and use mimics to analyze and process the data
- Supplement the data, and later use the ROC curve and calibration curve to verify the accuracy of the model, which is currently near 95%

Work Experience

Ant Group Co., Ltd., OceanBase SQL OPTIMIZER TEAM intern Jul. 2024 - Aug. 2024

- Assist open source oceanbase work of OceanBase, a distributed relational database which handle high-throughput and large-scale data processing in financial and e-commerce applications
- **Enhanced Performance and Efficiency:** Boosted UTF-8 validation speed by 20x-50x in high-production environments, optimized ASCII case batch computations and using SIMD instructions
- Expanded OceanBase's international character encoding support by implementing four East Asian charsets (e.g., ujis), enabling robust data processing for global users and markets
- **User Experience Enhancement:** Developed a virtual table feature for external table errors, allowing users to identify and resolve issues faster during external table construction

Step AI, Agent Development Intern Feb. 2024 - Jun. 2024

- **Built backend systems using trpc-go** to support the development of the RAG system for Step Chat, integrating PDF parsing, vector databases, and index building for enhanced AI-driven responses.
- **Designed and implemented an RBAC-based user control system** on the overall AI agent program, from database design, rbac structure design, to implementation. Improve the management of the whole agent system and structure
- **Developed internet-connected AI features**, enabling real-time access to news, weather, and other external data through API integrations with LLM
- **Led backend development for the OpenAPI platform**, collaborating with product and testing teams to deploy features in test and production environments. Managed virtual account for enterprise clients, data filtering, and automated synchronization to Feishu