Zhe Xu

https://phospheneser.github.io

Email: zxu24@m.fudan.edu.cn
GitHub: github.com/Phospheneser

EDUCATION

Fudan University — Shanghai, China
M.S. in Computer Science; Advisor: Prof. Xipeng Qiu — *Sep 2024 – Present*

Fudan University — Shanghai, China
B.S. in Artificial Intelligence, School of Computer Science — *Sep 2020 – Jun 2024*

EXPERIENCE

Moss Intelligence (SII) — Shanghai, China
Algorithm Research Intern — *Apr 2025 – Present*
- Large-scale speech-language model development and optimization.

NetEase Leihuo — Hangzhou, China
Algorithm Intern — *Jan 2024 – Mar 2024*
- Developed an NPC interaction demo driven by large language models, exploring AI applications in gaming.

Nuclear Power Institute of China, No. 8 Institute — Chengdu, China
Algorithm Intern — *Jul 2023 – Oct 2023*
- Participated in a clean-energy large model project, responsible for RLHF fine-tuning tasks.

Fudan University RM Team EGA — Shanghai, China
Co-founder & Vision Group Leader — *Oct 2024 – Dec 2024*
- Responsible for implementing the auto-aiming system including prediction and ballistic compensation.
- Team won the 1v1 Championship at Shanghai Regional and First Prize in 3v3 Battle Tournament.

PUBLICATIONS

- **DetectiveQA: Evaluating Long-Context Reasoning on Detective Novels** *ICLR 2025*
  - **Zhe Xu**, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, Xipeng Qiu
  - A narrative reasoning benchmark based on detective novels with very long contexts (avg. 100K+ tokens), offering 1,200 Chinese-English QA pairs with detailed reasoning chains.

- **VLABench: Language-Conditioned Robotics Manipulation Benchmark** *CVPR 2025*
  - Shiduo Zhang, **Zhe Xu**, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, Xipeng Qiu
  - First benchmark for VLA with long-horizon reasoning tasks, focusing on evaluation and definition.

- **MOSS-Speech: True Speech-to-Speech Models Without Text Guidance** *arXiv 2024*
  - SLM Team (contributor)
  - Modality-based layer splitting architecture with frozen pre-training strategy for native speech processing.

PROJECTS

- **LaCSR5-SpMV Optimization**: Optimized CSR5-based Sparse Matrix-Vector Multiplication for Loongson 3A5000, achieving up to $2.78\times$ speedup using SIMD and multi-core features. **Second Prize** in National Collegiate Computer Systems Ability Competition.
- **Personal Academic Homepage Template**: Open-source template built around configuration-first and modular design, supporting multilingual content and plugin extensions.
- **ESP32 Arduino Game Console**: Developed a game console using ESP32 and Arduino in C, including soldering, case design, and game programming.

SKILLS

- **Languages**: Python, C/C++, C#, SQL   **Research Areas**: NLP, Speech Models, Embodied AI

- **Technologies**: PyTorch, Transformers, RLHF, SIMD Optimization   **Tools**: Git, Linux, LaTeX