# Liming Lu

Email: liming_lu@sjtu.edu.cn | Tel: (86) 15868256485 | GitHub: Elemmire1.github.io

## Education

**Shanghai Jiao Tong University**, Shanghai, China — Sept. 2023 - Jun. 2027 (Expected)

**B.S. in Computer Science**, **John Hopcroft Class**

**GPA: 90.2/100 (Major), 90.3/100 (Overall)**, **Zhiyuan Honorary Scholarship**

**Zhiyuan Honors Program** (Top 10%)

## Publications

- **Liming Lu**, Kaixi Qiu, Jiayu Zhou, Jushi Kai, Haoyan Zhang, Huanyu Wang, Jingwen Leng, Ziwei He, Zhouhan Lin.
  One Size Does Not Fit All: Token-Wise Adaptive Compression for KV Cache
  ***ICML 2026*** *(Under Review).*
- Yunchong Song, Jushi Kai, **Liming Lu**, Kaixi Qiu, Zhouhan Lin.
  Towards Compressive and Scalable Recurrent Memory
  ***ICML 2026*** *(Under Review).*
- Haipeng Zhang, Yifan Liu, Xinyu Gu, **Liming Lu**, ... , Lei Bai, Zhouhan Lin, Ran Li
  One-for-all Intermittent Renewable Energy Forecasting with a Physics-guided Foundation Model
  ***Nature Energy*** *(Under Review).*

## Research Experiences

**One Size Does Not Fit All: Token-Wise Adaptive Compression for KV Cache**

Submitted to ***ICML 2026***, Advisor: Prof. Zhouhan Lin

- Addressed the performance degradation of dimensionality reduction KV compression methods under high compression ratios, which limits their practical effectiveness in resource-limited LLM inference.
- Proposed **DynaKV**, a token-wise adaptive KV cache compression method allowing the model to **dynamically allocate** different compression rates across tokens, which achieves more aggressive compression.
- Evaluated on both long- and short-context benchmarks, DynaKV significantly outperforms traditional methods such as PALU and MatryoshkaKV, maintaining stable accuracy even at **91%** compression ratios.

**Towards Compressive and Scalable Recurrent Memory**

Submitted to ***ICML 2026***, Advisor: Prof. Zhouhan Lin

- Addressed the bottleneck for scaling to long-context in Transformers, overcoming the trade-off between theoretical soundness and practical scalability in recurrent memory approaches.
- Proposed **Elastic Memory**, a HiPPO-grounded memory architecture that compresses historical sequences into a fixed-size state via **optimal online approximation**, with a flexible mechanism for efficient retrieval.
- Demonstrated consistent superiority on **32k+** long-context benchmarks across three domains, achieving up to **16×** memory efficiency over Memorizing Transformer, outperforming Melodi at all memory sizes (even with **30%** fewer parameters), and delivering significantly faster scaling at larger model sizes.

**One-for-all Intermittent Renewable Energy Forecasting with a Physics-guided Foundation Model**

Submitted to ***Nature Energy***, Advisor: Prof. Zhouhan Lin

- Tackled the poor predictability and limited generalization of existing IRES forecasting models across regions, weather conditions, forecasting horizons, and generation technologies.
- Designed **IresFM**, a foundation model for intermittent renewable energy forecasting, using a **two-stage training framework** with physics-guided large-scale pretraining on virtual sites and lightweight fine-tuning on limited real-world data to enable zero-shot generalization and local adaptation.
- Evaluated on real-world wind and solar farms, IresFM achieves **near–state-of-the-art** zero-shot performance and further reduces forecasting error by **10%** after fine-tuning.

## Skills

**Programming:** C++, Python, PyTorch, Transformers, Git, Linux, Hugging Face, LaTeX.

**Mathematics:** Linear Algebra, Mathematical Logic, Information Theory, Probability, Mathemetical Analysis.

**Language:** English - TOEFL: 106.