Xinglu **Wang**

Ph.D. Candidate | Simon Fraser University

(+1)-7783250568 | yanawlx@gmail.com | luzai | xinglu-wang-1734b2116 | Xinglu Wang

Education

Simon Fraser University
Burnaby, British Columbia

Computing Science, Ph.D. — *Sep. 2021 - April 2026*

- Cum. **GPA**: $4.13/4.33$.
- Helmut and Hugo Eppich Family Graduate Scholarship, Graduate Fellowships, Special Graduate Entrance Scholarship
- Supervised by **Prof. Jian Pei** and **Prof. Jiannan Wang**

Zhejiang University
HangZhou, Zhejiang

Information Engineering, M.S. — *Sep. 2018 - June 2021*

- Cum. **GPA**: $90.75/100$.
- Supervised **Prof. Yingming Li**, and **Prof. Zhongfei (Mark) Zhang**

Zhejiang University
HangZhou, Zhejiang

Information Engineering, B.Eng. — *Sep. 2014 - June 2018*

- Cum. **GPA**: $90.21/100$, $3.93/4$, **Ranking** $7^{th}/174$.
- Meritorious Winner, **Interdisciplinary Contest in Modeling** (ICM)
- **First**-Class Scholarship for Outstanding Students

Experience

Failure Recovery for Large-scale MoE Serving
Aug 2025–Feb 2026

Huawei Canada Research Institute, Burnaby — *Associate Researcher, Intern*

- Designed a failure-recovery mechanism for disaggregated MoE serving systems, where attention and MoE computation run on separate workers, enabling fast recovery without full instance restarts (*ReviveMoE, ArXiv 2026*). Implemented the core recovery path, including abstractions for attention and MoE executors and **proxy-based switching** for rapid role transitions.
- Developed a step-level recovery framework enabling safe rollback from unknown intermediate states via **transaction-log-based** recovery, including block-table restoration, request migration, and partial recomputation.

SLO-aware LLM Serving System
Dec 2024–Aug 2025

Huawei Canada Research Institute, Burnaby — *Associate Researcher, Intern*

- Worked on an LLM serving system supporting heterogeneous latency targets (*Service Level Objectives, SLOs*), designing heuristic **SLO-aware** dispatching and scaling algorithms to balance user latency requirements and serving cost. Published on *ArXiv 2025*.
- Gained hands-on experience with **fast-scaling** mechanisms for responsive elastic scaling, including pre-started serving processes, device-to-device weight transfer, and stage-aware orchestration under disaggregation mode.

Disaggregated Large Multimodal Model Serving
Feb 2024–Dec 2024

Huawei Canada Research Institute, Burnaby — *Associate Researcher, Intern*

- Worked with the **Encode–Prefill–Decode (EPD)** disaggregation architecture for serving large multimodal models, published in **ICML 2025**. Supported *experimental evaluation* of the serving system, analyzing efficiency and limitations under different scenarios.
- Became familiar with the system design, including inter-stage coordination and *cache layout design*/management, and gained exposure to system techniques such as peer-to-peer device communication, **CUDA IPC** memory sharing, and CUDA virtual memory management.

LLM for Mathematical Modeling
Feb 2023–Feb 2024

Huawei Canada Research Institute, Burnaby — *Associate Researcher, Intern*

- Evaluated LLM capabilities, with prompt engineering and supervised fine-tuning (SFT), for mathematical modeling across optimization problems of varying complexity. Co-authored a survey in **EJOR 2025** on how AI can enhance multiple stages of the optimization pipeline.
- Proposed a graph-edit-distance-based evaluation metric for linear programming word problems that better aligns with human judgment and identifies mathematically equivalent formulations; published at **COLING 2024**.

Smart Initial Basis Selection for Linear Programs

Feb 2022–Feb 2023

Huawei Canada Research Institute, Burnaby — *Associate Researcher, Intern*

- Proposed a *graph neural network* approach to predict **close-to-optimal** initial bases for linear programming solvers.
- Combined model prediction with post-processing steps to generate *valid bases* for simplex warm starts.
- Demonstrated substantial speedup over rule-based baselines across multiple datasets and solvers. Work published in **ICML 2023**.

Harmonized Multi-exit Learning

Dec 2019–Aug 2021

Data Science & Engineering Research Center, ZJU — *Master Thesis*

- Studied multi-exit learning for adaptive inference, addressing optimization interference between exits.
- Proposed *gradient deconfliction training* using gradient projection to resolve exit conflicts; published in **IEEE ICIP 2020**.
- Designed a *meta-learning*-based *harmonized weighting* scheme for dense teacher-student distillation across exits, accepted at **AAAI 2021**.
- Evaluated on large-scale ImageNet using *Cloud TPU*; reported a bug in **pytorch/xla**.

Person Re-identification

Oct 2017–June 2018

Data Science & Engineering Research Center, ZJU — *Undergraduate Thesis*

- Analyzed the pipeline (data sampling, feature extraction, loss design, post-processing) and summarized in a *technical blog*.
- Open-source contributions: (1) proposed a Cython module in **deep-person-reid** to accelerate evaluation by **20×**; (2) fixed ResNet depth bug in **pytorch-classification** enabling fair CV model comparison.

Publications

[1] Li, H., **Wang, X.**, Feng, C., Zuo, C., Wang, Y., Lo, H., Cui, Y., Wang, B., Cui, D., Jing, S., Shan, Y., Xiong, Y., Wang, J., Zhang, Y., and Fan, Z. 2026. ReviveMoE: Fast recovery for hardware failures in large-scale MoE LLM inference deployments. arXiv preprint arXiv:2602.21140.

[2] Yousefijamarani, Z., **Wang, X.**, Wang, Q., Heisler, M., Shabani, T., Gholipour, N., Yassini, P., Chang, H., Chen, K., Zhang, Q., Bai, X., Wang, J., Xiong, Y., Zhang, Y., and Fan, Z. 2025. HyperFlexis: Joint design of algorithms and systems for multi-SLO serving and fast scaling. arXiv preprint arXiv:2508.15919.

[3] Fan, Z., Ghaddar, B., **Wang, X.**, Xing, L., Zhang, Y., and Zhou, Z. 2025. Artificial intelligence for optimization: Unleashing the potential of parameter generation, model formulation, and solution methods. *European Journal of Operational Research*.

[4] Singh, G., **Wang, X.**, Hu, Y., Yu, T., Xing, L., Jiang, W., Wang, Z., Bai, X., Li, Y., Xiong, Y., Zhang, Y., and Fan, Z. 2025. Efficiently serving large multimodal models using EPD disaggregation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

[5] Heisler, M., Yousefijamarani, Z., **Wang, X.**, Wang, Q., et al. 2025. LLM inference scheduling: A survey of techniques, frameworks, and trade-offs. *TechRxiv Preprint*. DOI: 10.36227/techrxiv.176238087.79673350.

[6] Xing, L., **Wang, X.**, Feng, Y., Fan, Z., Xiong, J., Guo, Z., Fu, X., Ramamonjison, R., Mostajabdaveh, M., Han, X., Zhou, Z., and Zhang, Y. 2024. Towards human-aligned evaluation for linear programming word problems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (COLING)* (pp. 16550–16556).

[7] Fan, Z., Huang, F., **Wang, X.**, Zhou, Z., Pei, J., Friedlander, M. P., and Zhang, Y. 2024. Fair and efficient contribution valuation for vertical federated learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[8] Fan, Z., **Wang, X.**, Yakovenko, O., Sivas, A. A., Ren, O., Zhang, Y., and Zhou, Z. 2023. Smart initial basis selection for linear programs. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* (pp. 9650–9664).

[9] Gholami, M., Akbari, M., **Wang, X.**, Kamranian, B., and Zhang, Y. 2023. ETran: Energy-based transferability estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 18613–18622).

[10] Qiao, C., Xiang, Z., **Wang, X.**, Chen, S., Fan, Y., and Zhao, X. 2023. Objects matter: Learning object relation graph for robust absolute pose regression. *Neurocomputing*, 521, 11–26.

[11] **Wang, X.**, and Li, Y. 2021. Harmonized dense knowledge distillation training for multi-exit architectures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11), 10218–10226.

[12] Ouyang, S., **Wang, X.**, Lyu, K., and Li, Y. 2021. Pseudo-label generation-evaluation framework for cross-domain weakly supervised object detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 724–728).

[13] **Wang, X.**, and Li, Y. 2020. Gradient deconfliction-based training for multi-exit architectures. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 1866–1870).