Junfei Zhan

+86 18929985977 / +1 2679219547 zjf2024@seas.upenn.edu junfei-z.github.io

Education

University of Pennsylvania **Aug. 2024 – *Expected* May 2026**
MS in Electrical Engineering *Philadelphia, USA*

GPA: 4.00/4.00

University of Birmingham & Jinan University **Sep. 2020 – June 2024**
BSc in Applied Mathematics with Information Computing Science *Birmingham, UK*
Dual Degree - BSc in Information and Computing Science *Guangzhou, CN*

GPA: 3.82/4.00 (Birmingham) & 4.01/5.00 (Jinan) **Honors:** First Class (Birmingham)

Relevant Coursework

- Autonomous Driving (A+)
- Graph Neural Networks (A)
- Applied Machine Learning (A)
- Information Theory (A)
- Linear System Theory (A)
- MATLAB Programming (A+)
- Game Theory (A+)
- Applied Statistics (A)
- Convex Optimization (A+)

Honors & Awards

President Gutmann Leadership Award, University of Pennsylvania, 2025

Professional Student Travel Grant, University of Pennsylvania, 2025

Outstanding Undergraduate Thesis Award, Jinan University, 2024

First-class Scholarship, Jinan University, 2023 & 2024

China National Scholarship (Top 0.2% – Highest national honor for Chinese undergraduates), 2022

Publications

Junfei Zhan, Jiayi Wu, Tengjiao He, Kwan-Wu Chin, "Task Offloading and Approximate Computing in Solar Powered IoT Networks," *IEEE Networking Letters*, vol. 6, no. 1, pp. 26-30, March 2024, doi: 10.1109/LNET.2023.3328893.

Junfei Zhan, Haoxun Shen, Zheng Lin, Tengjiao He, "PRISM: Privacy-Aware Routing for Adaptive Cloud–Edge LLM Inference via Semantic Sketch Collaboration." *AAAI 2026* (Accepted, CCF-A).

Haoxun Shen, **Junfei Zhan**, Tengjiao He, "RL-Enhanced Disturbance-Aware MPC for Fast and Robust UAV Trajectory Tracking." *IEEE SMC 2025* (Accepted, Tsinghua-A).

Junfei Zhan, Weihang Ding, Haoxun Shen, Tengjiao He, "Stochastic Power Modeling and Constrained MDP Optimization for On-Device SLM Inference," *ICASSP 2026 Conference* (Under Review, CCF-B).

Junfei Zhan, Tengjiao He, Kwan-Wu Chin, Fei Song, "Orchestrating Data Collection and Computation in Green Virtualized Networks." *IEEE Internet of Things Journal* (Under Review, SCI Q1).

Research Projects

Stochastic Power Modeling and Constrained MDP Optimization for On-Device SLM Inference

Lead Author, Advisors: Prof. Tengjiao He ***Jul. 2025 – Sep. 2025***

- Built Hidden Semi-Markov Models (HSMMs) to capture stochastic CPU/GPU power dynamics and phase durations during on-device small language model (SLM) inference.
- Integrated HSMM-based energy estimates with LLM-as-Judge scoring to model the trade-off between inference quality and energy consumption in on-device SLM inference.
- Formulated a constrained Markov decision process (CMDP) with Q-learning to adaptively schedule inference tasks under finite energy budgets and device-level power caps.

Privacy-Aware Routing for Adaptive Cloud–Edge LLM Inference via Semantic Sketch Collaboration

Lead Author, Advisor: Prof. Tengjiao He ***Mar. 2025 – Aug. 2025***

- Proposed PRISM, a privacy-aware framework that dynamically routes user prompts between cloud, edge, and collaborative modes based on entity-level sensitivity.
- Designed an adaptive two-layer local differential privacy mechanism and a semantic sketch pipeline to balance privacy, utility, and system latency.
- Deployed PRISM on real-world cloud-edge infrastructure, achieving 65% lower energy consumption and 1.54× faster inference over baseline privacy-preserving methods.

Orchestrating Data Collection and Computation in Virtualized Networks
Research Assistant, Advisor: Prof. Kwan-Wu Chin ***May 2024 – Dec. 2024***

- Formulated a mixed integer linear programming (MILP) model to optimize the embedding and scheduling of Virtual Network Functions (VNFs) in solar-powered IoT networks, minimizing the maximum age of service (AoS).
- Developed a Receding Horizon Control Optimization (RHCOP) algorithm leveraging Gaussian Mixture Models (GMM) to predict energy arrivals and wireless channel gains for efficient resource allocation.
- Implemented resource-aware embedding of Directed Acyclic Graph (DAG) requests to balance computational loads, manage energy constraints, and enhance the freshness of IoT services.

Competitions

Kaggle: Google - Fast or Slow? Predict AI Model Runtime ***Oct. 2023***
Prize: Bronze Medal (Top 10%)

- Preprocessed data and built a Graph Convolutional Network (GCN) to predict runtime of graphs and configurations using a machine learning model trained on runtime data.

2022 China Undergraduate Mathematical Contest in Modeling ***Sep. 2022***
Prize: Provincial First Prize (Top 10%)

- Applied Logistic Regression, Fuzzy C, Principal Component Analysis, and K-Means Clustering to conduct weathering analysis and classification of ancient glass artifacts.

2022 Mathematical Contest in Modeling ***Feb. 2022***
Prize: Honorable Mention (Top 30%)

- Utilized Empirical Mode Decomposition (EMD) and Long Short-Term Memory (LSTM) neural networks to develop a portfolio strategy for gold and bitcoin.

Professional Experience

Graduate Teaching Assistant **Sep. 2025 – Present**
University of Pennsylvania *Philadelphia, PA*

- Designed homework assignments for around 125 students in *ESE 5000 Linear System Theory.*
- Held office hours to provide academic support and address individual student needs.
- Led recitations to reinforce lecture material, facilitate problem-solving, and manage Q&A in a large-class setting.

JLL (Jones Lang LaSalle) **Aug. 2023 – Oct. 2023**
Intern in Data Analytics *Shenzhen, China*

- Collected and analyzed data, including tenants, vacancy rates, and business performance metrics, through on-site surveys of 600+ companies and 40+ Grade-A office buildings.
- Utilized SQL for data extraction and Excel for updating databases, ensuring accurate and up-to-date information.
- Developed a program to extract text from large image datasets and efficiently manage text files, significantly improving workflow efficiency.

Technical Skills

Programming Languages: C, Python, SQL, MATLAB, R, SPSS
Algorithms: PID Controller, Dijkstra, Dynamic Programming, Receding Horizon Control, Simplex, Big-M Method
Frameworks: PyTorch, Pyomo, scikit-learn