Wangshu Zhu

☎ (+1) 646-240-9849 ✉ wz2708@columbia.edu 🌐 wz2708.github.io github.com/wz2708

Education

Columbia University **Sep. 2024 – May 2026**

Master of Science in Electrical Engineering *New York, US*

- Awards: 2025 Spring MS Honors Students
- Courses: Deep Learning & Neural Network, Reinforcement Learning, Generative AI, Advanced Big Data & AI

Wuhan University **Sep. 2020 – May 2024**

Bachelor of Science in Information Engineering *Wuhan, China*

Experence

FairTraj-COSMOS: Trajectory Prediction under Missingness Protocols **May 2025 – August 2025**

Summer Research Assistant, AIDL Lab, Columbia University *New Tork*

- Identified a critical overlooked bias in trajectory prediction benchmarks: model performance often affected by how missing tracks are reconstructed rather than the predictive capacity of the model design.
- Curated and processed a long-term NYC traffic dataset (COSMOS), extracting trajectories from raw video, converting them into a nuScenes-compatible format, and designing four principled missingness protocols for systematic evaluation.
- Standardized experimental design by employing Unitraj as a unified training spine, enabling controlled comparisons of AutoBot, Wayformer, and MTR, and incorporating Brier-FDE as a calibration-aware complement to FDE.
- Delivered quantitative evidence that Protocol of Train-Rich & Eval-Pure consistently yields the best accuracy–calibration trade-off, showing how protocol choice can overturn "state-of-the-art" claims.

Dual-RAG Multi-Agent Clinical Decision Support System **March 2025 – June 2025**

Data Scientist Internship at Graphen, Inc *New York*

- Tackled real-world diagnostic uncertainty from free-text patient narratives, where single-agent LLMs often misclassify due to reasoning shortcuts and lack of evidence grounding, motivating a multi-agent, knowledge-aware framework.
- Designed a clinical workflow–inspired pipeline: Reception Doctor extracts structured symptoms, General Doctor proposes candidates via Dual-RAG, and Critic–Expert debates iteratively validate lever symptoms, reducing diagnostic ambiguity
- Advanced retrieval with Dual-RAG, integrating a UMLS-backed biomedical knowledge graph with a continually self-updating case memory, where each new case is where each diagnosed case is agent-processed and re-ingested.
- Achieved 96.5% Top-1 accuracy and 0.98 MRR on held-out test sets, surpassing single-agent (73.9%) and CoT baselines (77.1%) while maintaining interpretability; submitted to AMIA 2025 Late-Breaking Paper track.

Interpretable Multimodal Framework for Collision Prediction **January 2025 – May 2025**

Research Assistant, DitecT Lab, Columbia University *New York*

- Addressed the poor reliability and explainability for driver assistance systems in extreme scenarios (low-light, occluded, or noisy scenes) by targeting early collision forecasting from dashcam video.
- Curated multimodal cues by combining object detections, segmentation masks, depth maps, and optical flow, and standardized them into a unified representation of scene dynamics to overcome modality fragmentation.
- Introduced a novel conversion layer that reformulates heterogeneous visual dynamics into dialogue-style prompts, enabling a large vision–language model (Qwen2.5-VL) to perform temporal collision forecasting with rationales.
- Delivered evidence that the VLM-based approach achieves 0.69/0.61 mAP at 0.5s/1.0s horizons on Nexar Dashcam, showing that our method yields both stronger early-warning and calibrated, auditable outputs for safety-critical AI.

RoPE Integration in Performer for Vision Transformer **October 2025 – December 2025**

Graduate Course Research Project *New York*

- Addressed the gap between linear attention efficiency and spatial modeling fidelity by embedding Rotary Positional Embeddings (RoPE) into the Performer, aiming to preserve both O(N) scalability and 2D positional expressiveness.
- Developed a unified 2D RoPE–Performer module by extending rotary embeddings into axial and mixed variants and embedding them into Performer's query–key projections, enhancing spatial sensitivity while retaining O(N) efficiency.
- Improved Top-1 accuracy by +1.8 percentage points and reduced GPU memory consumption by 13% on ImageNet-1K, demonstrating that RoPE–Performer can serve as an effective building block for scalable vision and foundation models.

Publication

- Wangshu Zhu. Dual-RAG with Conditional Reviewer: A Multi-Agent Framework for Clinical Decision Support. Submitted to AMIA 2025 Annual Symposium (Late-Breaking Paper Track, 2025).
- Wangshu Zhu. The Application of Augmented Reality Technology in SDA Patients. 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (2023).

Technical Skills

Python, Pytorch, Tensorflow, Java, SQL, and Overleaf