

JUNLI GONG

✉ gong.jun@northeastern.edu ✉ gongjunli0307@gmail.com 📞 +1 206-741-8484 📞 +86 152-2086-9984
🌐 junli-gong.github.io 🌐 github.com/junli-gong 🎓 Google Scholar

RESEARCH INTERESTS

Efficient, scalable systems for multimodal and mixture-of-experts models, with a focus on LLM inference, communication-aware distributed inference, and hardware-aware acceleration.

EDUCATION

Northeastern University

Sep 2025 – Dec 2026 (expected)

Master of Science in Computer Science, Khoury College of Computer Sciences

Seattle, WA, USA

- **GPA:** 3.89/4.00; **Seattle Campus Scholarship** (merit-based)
- **Relevant Courses:** Machine Learning, Scalable Distributed Systems, Reinforcement Learning, Algorithms

Hong Kong Baptist University

Sep 2021 – Jun 2025

Bachelor of Science (Honours) in Computer Science and Technology

Zhuhai, China

- **Honors:** Second-Class Scholarship (AY 2024–25); President’s Honour Roll; Dean’s List (4 semesters)

PUBLICATIONS

* denotes equal contribution.

Peer-reviewed

- [1] Guo, Y., **Gong, J.**, Cai, H., Cheung, Y.-m., & Su, W. (2026). Can Segmentation Models Understand the World? Proactive Affordance Reasoning via Vision-Grounded Chain-of-Thought. *EMNLP 2026* (Main Conference, **Oral**).
- [2] Guo, Y.*, **Gong, J.***, Wang, W., Cai, H., Cheung, Y.-m., & Su, W. (2026). PEAM: Parametric Embodied Agent Memory for Continual Skill Learning via Internalization of Experience in Minecraft. *EMNLP 2026* (Findings, Poster).
- [3] Guo, Y., **Gong, J.**, Dong, W., Cheung, Y.-m., & Su, W. (2026). Adding Thermal Awareness to Visual Systems in Real-Time via Distilled Diffusion Models. *CVPR 2026 Workshop* (World Models Meet Active Sensing, Poster).
- [4] Zheng, C.*, **Gong, J.***, Guo, J., Tang, Z., & Wang, T. (2025). Enhancing Federated Learning in IoT Through Dynamic Spectral Clustering. *WASA 2025* (**Oral**).
- [5] Fei, C., Xia, N., Tsai, P.-W., Lu, Y., Pan, X., & **Gong, J.** (2024). An Effective Feature Selection Algorithm for Machine Learning-based Malicious Traffic Detection. *AsiaJCIS 2024*.

Under Review

- [6] Guo, Y.*, **Gong, J.***, Lu, Y., Xu, X., Cheung, Y.-m., & Su, W. (2026). Bringing Multimodal Large Language Models to Infrared-Visible Image Fusion Quality Assessment. Preprint. arXiv:2605.06969.
- [7] Guo, Y., **Gong, J.**, Cheung, Y.-m., & Su, W. (2026). LumiVideo: An Intelligent Agentic System for Video Color Grading. Preprint. arXiv:2604.02409.
- [8] Guo, Y., **Gong, J.**, Cheung, Y.-m., & Su, W. (2024). Fuse4Seg: Image Fusion for Multi-Modal Medical Segmentation via Bi-level Optimization. Preprint. arXiv:2409.10328.

RESEARCH EXPERIENCE

Graduate Researcher

Jan 2026 – Present

Khoury College of Computer Sciences, Northeastern University

Advisor: Prof. Yifan Hu

- Investigating speculative expert pre-dispatch for Mixture-of-Experts inference: lightweight linear probes predict expert routing from pre-attention hidden states so that the expert-parallel AllToAll can overlap with attention, while a verify-and-fallback step keeps outputs identical to the serial path.
- On DeepSeek-V2-Lite (prefill, 2×A100, EP=2) with a 5 ms per-layer AllToAll delay injected to emulate PCIe/cross-node interconnects, the SGLang prototype reduces median and P99 time-to-first-token by 22%.

Research Assistant

Sep 2024 – Present

Department of Computer Science, Beijing Normal-Hong Kong Baptist University

Supervisor: Prof. Weifeng Su

- Drove the experimental research for six papers on multimodal perception and embodied AI [1–3, 6–8]: built the training and evaluation pipelines, designed and ran the ablation studies, and wrote the experiments sections.
- Co-first-authored PEAM [2] (Findings of EMNLP 2026): designed the failure-correction contrastive training objective that internalizes corrected trajectories into per-category MoE-LoRA skill adapters, and implemented and evaluated the full framework in Minecraft, reaching 69.7% success on a held-out long-horizon task suite—+15.2 percentage points over VOYAGER and +9.1 over Optimus-1—with 6.8× fewer tokens per task.
- Co-first-authored FuScore [6]: designed the Thurstone-based within- and cross-source ranking objectives and built the MLLM-based infrared-visible fusion quality assessor around them, from zero-shot scoring to fine-tuning with soft-label distributional supervision; it outperforms every external baseline on all five rank metrics on EVA Fusion (+0.094 SRCC over the strongest), with the soft-label/tripartite-loss recipe alone adding +0.111 SRCC on the same backbone.

- For SegWorld [1] (EMNLP 2026), implemented the two-pass reasoning pipeline (instruction-free proactive observation, then instruction-conditioned visual chain-of-thought) on Sa2VA-Qwen2.5-VL-7B and ran the Intent2Part evaluation (intent-level mIoU 0.514 zero-shot, 0.612 with intent supervision, above all fine-tuned baselines); for FusionProxy [3], built the distilled-diffusion real-time thermal-visible fusion module (84 FPS on an RTX 4070 in FP16) and its closed-loop CARLA driving evaluation.

Research Intern

Jul 2025 – Aug 2025

Pazhou Lab / School of Computer Science and Engineering, South China University of Technology

Supervisor: Prof. Lu Lu

- Developed a custom FP16 Tensor Core HGEMM kernel in CUDA for Transformer-shaped workloads on NVIDIA A800, with coalesced global-memory access and double-buffered software pipelining (+6–9% over the same kernel without pipelining).
- Benchmarked against cuBLAS on five BERT/LLaMA-derived shapes: $1.33\times$ faster on the bandwidth-bound attention-score shape ($1024 \times 1024 \times 64$); on the four compute-bound shapes, where the kernel reached 11–39% of cuBLAS, profiling identified shared-memory bandwidth and pipeline depth as the bottlenecks, motivating CUTLASS-style multi-stage pipelining (overlapping shared-memory-to-register loads with MMA execution) and shape-adaptive kernel selection as the next steps.

Undergraduate Researcher

Jan 2024 – Jun 2025

Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai

Supervisor: Prof. Jianxiong Guo

- Co-first-authored a WASA 2025 paper [4] on clustered federated learning for non-IID IoT data: designed DSCFL, a dynamic spectral-clustering scheme that regroups clients by data similarity during training, and implemented it in Python together with a dynamic resource-allocation strategy.
- DSCFL beats FedAvg and CFL (Sattler et al., TNNLS’20) on EMNIST, CIFAR-10, and FashionMNIST in all six settings—up to +7.9 percentage points over FedAvg and +6.4 over CFL with 80 clients—and recovers the correct client clusters by round 5 versus round 62 for CFL.

Research Assistant

Jan 2024 – Jun 2024

School of Computer and Electronic Information, Nanjing Normal University (remote)

Supervisor: Dr. Nian Xia

- Collected and preprocessed traffic datasets (TII-SSRC-23, UNSW-NB15), trained SVM and decision-tree detectors with EFS-CST feature selection, and wrote the data-collection and evaluation sections of the AsiaJCIS 2024 paper [5].

INDUSTRY EXPERIENCE

iFLYTEK Co., Ltd.

Jul 2026 – Aug 2026

AI Infra Intern

Hefei, China

- Developed and debugged an automated evaluation platform for adapting speech models (ASR/TTS) to heterogeneous domestic AI accelerators (Ascend NPU, DCU, MLU, MUSA): resolved 30 engineering issues, redesigned task-liveness detection, and built a result-integrity validator that flagged all 99 historical faulty records with zero false positives.
- Root-caused a WER degradation of Parakeet-CTC-1.1B on Ascend NPU via layer-wise operator comparison (1,811 ops): an NPU-only branch in Hugging Face Transformers cast the float relative-position-bias mask to boolean, discarding positional information across 42 attention layers; fix verified on a 200-sample regression and reported upstream.
- Led a cross-platform accuracy-drift audit against GPU baselines, separating engineering artifacts from genuine platform drift across 8 outlier models and establishing comparability pre-checks (model ID, weights, inference engine, decoding parameters) for chip-accuracy statistics.

Shenzhen Quant-Cloud Energy Network Technology Co., Ltd.

Aug 2024

Research & Development Intern, Artificial Intelligence Department

Shenzhen, China

- Built an Anomaly Transformer-based bearing-temperature anomaly detector for wind-turbine gearbox predictive maintenance.

Kenexs Intelligent Technology Co., Ltd.

Jul 2023 – Sep 2023

Software Development Intern

Shenzhen, China

- Built a Qt log analyzer for industrial machine-vision cameras used by the test team.

SKILLS

Programming: Python, C/C++, CUDA, Java, SQL, JavaScript
ML & Systems: PyTorch, Hugging Face Transformers, SGLang, NeMo, NumPy/SciPy, CUDA kernel optimization (Tensor Cores, cuBLAS), Linux, Git, LaTeX
Languages: Chinese (native), English (fluent)