

Juhyung Lee

[\[LinkedIn\]](#) [\[Github\]](#) [\[Google Scholar\]](#) [\[Website\]](#)

Email: juhyung.lee@outlook.com

Mobile: +1-213-245-9356

Santa Clara, CA, USA

SUMMARY

- **Applied AI Scientist (PhD)** specializing in **Multimodal Foundation Models, On-Device Agentic Systems** for real-time reasoning and decision-making systems.
- Designed and deployed large foundation models (Llama3, Phi4, QWEN3) using **quantization, distillation (GPTQ, AWQ)** and **post-training (SFT, DPO, ORPO)** for real-time, on-device deployment on **Apple Silicon (iPad/MacBook)**.
- Track record (**NeurIPS, ICML**): agentic LLM controllers; digital twins for synthetic data generation and evaluation; generative model compression.

PROJECT HIGHLIGHTS

- **On-Device Agentic Intelligence (NeurIPS'25 [1])**—Architected a real-time multi-agent system (“PEARL”) for cross-layer optimization; achieved $< 20\text{ms}$ inference latency on mobile GPUs via 4-bit quantization and LoRA adapters, demonstrating feasibility for interactive edge applications.
- **Multimodal Context Awareness (ICML'25 [2])**—Developed a context-aware policy engine fusing location, time, and RF sensor data; utilized SFT and DPO to align LLM reasoning for robust decision-making in dynamic physical environments.
- **Synthetic Data & Digital Twins (ICML'25 [3])**—Built an ML-driven digital twin with $\sim 1\text{ms}$ inference latency for rapid sim-to-real evaluation and large-scale synthetic data generation.
- **Generative compression [4]**—VQ-VAE/Diffusion for high-efficiency compression of high-dimensional sensor state; $8\times$ bandwidth reduction suitable for multimodal data streams.

EXPERIENCE

- **Nokia** Sunnyvale, CA, USA
Principal Researcher, AI/ML Aug. 2024 – Present
 - Led the development of an on-device agentic controller capable of system optimization. Integrated Llama3/Phi4 with real-time sensor streams. Optimized for **Apple iPad Pro** using GPTQ/AWQ quantization, achieving real-time inference constraints. [1] [\[Github\]](#) [\[Demo\]](#).
 - Designed a context-aware roaming agent that fuses multimodal inputs (geolocation, time-series signal quality) to predict actions. Fine-tuned foundation models using SFT, DPO, ORPO to align agent behavior with user connectivity goals [2] [\[Github\]](#) [\[Demo\]](#).
 - Built a RAG pipeline for technical specification analysis, enhancing engineering productivity by indexing complex unstructured data using vector databases and semantic search.
 - Contributor to Wi-Fi Standard (IEEE 802.11 AIML TIG/SC, TGbn), defining protocols for AI-native interfaces [5].
- **University of Southern California** Los Angeles, CA, USA
Postdoctoral Researcher, Wireless Devices and Systems Group (Head: [Prof. Andreas Molisch](#)) Apr. 2022 – Aug. 2024
 - **Digital Twins & Sim-to-Real:** Engineered a **ray-tracing** framework to generate synthetic datasets for **Physical AI**, overcoming data scarcity; validated on real system with $\approx 1\text{ms}$ latency and RMSE -14dB [3, 6].
 - **Neural Signal Tokenization:** Integrated Transformer-based (BART) models with **NVIDIA Sionna**; achieved 50% data reduction with high fidelity for **multimodal transmission** [7].
- **Samsung Research America** Dallas, USA
Senior AI/Wireless Research Engineer Dec. 2023 – Jan. 2024
 - **Generative Diffusion for State Estimation:** Applied **Latent Diffusion Models** and VQ-VAE to model high-dimensional sensor states. Achieved $8\times$ compression with high reconstruction fidelity [4] [\[Github\]](#).
- **Korea University** Seoul, Korea
Research Professor, Research Institute for Information & Communication Sep. 2021 – Feb. 2023
 - **Distributed Multi-Agent Control:** Reframed low-level system protocols (PHY/MAC) as decentralized agentic coordination tasks; developed Multi-Agent RL policies for high-mobility environments (3GPP TR-38.821) [8].

SKILLS

- **Frameworks & Code:** PyTorch, TensorFlow, Swift, Python, C/C++, CUDA, FSDP; NVIDIA Sionna, WirelessInSite.
- **Generative AI & Multimodal:** LLMs (Llama3, Phi4), Transformers (BART), Diffusion Models, VQ-VAE.
- **RL & Agentic AI:** Post-Train (RLHF/DPO/ORPO), Agentic Pipelines (task decomposition, planning, multi-step reasoning), RAG, Tokenization.
- **On-Device & Efficient ML:** Quantization (GPTQ, AWQ), Distillation, LoRA; Apple CoreML, latency profiling.

EDUCATION

- **Korea University** Seoul, Korea
Ph.D. in Electrical and Computer Eng. (Awarded by Research Excellence) Mar. 2016 – Aug. 2021
- **Korea University** Seoul, Korea
B. Eng. in Electrical and Electronic Eng. (National Sci. & Tech. Scholarship) Mar. 2011 – Feb. 2016

HONORS AND AWARDS

- [Program Committee](#), AI4NextG @*NeurIPS'25*
- [Industry Panelist](#), ML4Wireless @*ICML'25*
- [1st Place](#), Signal Processing Grand Challenges, @*ICASSP'23* [9]
- **Best Paper** @*ICTC'22*, **Best Paper** @*ICTC'21*
- **Grand Prize**, *Graduate Research Excellence Award* @Korea Univ.

SELECTED STANDARDS & PATENTS

- **IEEE 802.11 (AIML/TGbn):**
 - [IEEE 802.11] “On-device AI for context-aware WLAN roaming control,” *IEEE 802.11 AIML SC doc 25/2046* [5]
 - [IEEE 802.11] “On The Switching Criteria For Non-Primary Channel Access,” *IEEE 802.11 TGbn doc 24/1800*
 - [IEEE 802.11] “CSI in Sounding in Coordinated Beamforming,” *IEEE 802.11 TGbn doc 24/1823*
- **US Patents:**
 - [Patent-USA] “Deep reinforcement learning-based random access method for low earth orbit satellite network and terminal for the operation”, *US20230189353A1*, 2023
 - [Patent-USA] “Apparatus based on wireless optical communication”, *US20230083544A1*, 2022

SELECTED PUBLICATIONS [[LINK FOR FULL-LIST](#)]

- [1] J.-H. Lee*, Y. Lu*, and K. Doppler, “PEARL: Peer-enhanced adaptive radio via on-device LLM,” *NeurIPS*, 2025. [[demo](#)].
- [2] J.-H. Lee*, Y. Lu, and K. Doppler, “On-device LLM for context-aware Wi-Fi roaming,” *ICML*, 2025. [[paper](#)] [[code](#)] [[demo](#)].
- [3] J.-H. Lee* and A. F. Molisch, “AutoBS: Autonomous base station deployment with reinforcement learning and digital network twins,” *ICML*, 2025. [[paper](#)] [[code](#)].
- [4] J.-H. Lee*, J. Lee, and A. F. Molisch, “Generative vs. predictive models in massive MIMO channel prediction,” *Asilomar Conf. on Signals, Systems, and Computers*, 2024. [[paper](#)] [[code](#)].
- [5] J.-H. Lee* and et al., “On-device AI for context-aware WLAN roaming control at the STA,” *IEEE 802.11 AIML SC doc 25/2046*. [[IEEE Standard](#)] [[slide](#)].
- [6] J.-H. Lee* and A. F. Molisch, “A scalable and generalizable pathloss map prediction,” *IEEE TWC*, 2024. [[paper](#)] [[code](#)].
- [7] J.-H. Lee*, D.-H. Lee, J. Lee, and J. Pujara, “Integrating pre-trained language model with physical layer communications,” *IEEE TWC*, 2024. [[paper](#)] [[code](#)].
- [8] J.-H. Lee*, A. F. Molisch, and et al., “Handover protocol learning for LEO satellite networks: Access delay and collision minimization,” *IEEE TWC*, 2024. [[paper](#)].
- [9] J.-H. Lee*, A. F. Molisch, and et al., “PMNet: Large-scale channel prediction system for radio map prediction challenge,” in *IEEE ICASSP*, 2023. [[1st-Rank in ML Competition](#)] [[code](#)].