

Kevin Galim

Senior AI Research Engineer | Efficient LLM Inference, Post-Training, AI Systems
galimkevin@gmail.com · kevingalim.com · scholar.google.com · linkedin.com/in/kgalim/

Summary

Machine learning researcher working on efficient LLM inference, post-training systems, and accelerator-aware generative model pipelines. First/co-first author of publications at ICLR, ICML, ACL, ECCV, and WACV, with work spanning KV-cache and prompt/context compression, diffusion LLM parallel decoding, and parameter-efficient adaptation for state space models.

Experience

Senior AI Research Engineer

FuriosaAI

Jun 2021 – Present

Seoul, South Korea

- Conducted research on large-scale generative models and LLMs, including efficient inference, KV-cache optimization, diffusion LLMs, PEFT, state space models, and advanced architectures.
- Studied post-training systems including asynchronous OPD/RL-style pipelines, stale rollout effects, teacher-cache constraints, and throughput-quality trade-offs.
- Co-authored multiple first- and co-first-author papers in top-tier conferences: ICLR, ICML, ACL, CVPR, ECCV, and WACV.
- Designed end-to-end pipelines for training and evaluating LLMs, including accelerator-aware rollout generation and inference pipelines on custom AI hardware.

Technologies: PyTorch, Python, vLLM, FSDP, Custom AI Accelerators

AI/Computer vision research and development

Funzin

Mar 2020 – Jun 2021

Seoul, South Korea

- Developed computer-vision models for object detection, segmentation, gesture detection, and autonomous golf cart perception.
- Optimized models for embedded deployment using TensorRT, OpenVINO, Coral, DSP acceleration, and ARM NEON.
- Built and demonstrated a real-time OpenGL 3D surround-view system at CES 2021.

Technologies: PyTorch, TensorFlow, C++, TensorRT, OpenVINO, ARM NEON

Web / AR Developer

Freelance

Aug 2019 – Mar 2020

Munich

- Built AR mobile apps, cloud-backed web apps, and cross-platform mobile apps using Unity3D, Flutter, AWS, and Google Cloud.

Technologies: Unity3D, Flutter, AWS, Google Cloud

C++/CUDA software engineer

ARRI

Aug 2015 – Sep 2016

Munich

- Developed GPU-accelerated image-processing algorithms using CUDA and OpenCL.
- Built C++/OpenGL visualization and image-analysis tools for digital cinema workflows.

Technologies: C++, CUDA, OpenGL, OpenCL

Ongoing Work

- **AsyncOPD: How Stale Can On-Policy Distillation Be?**. *Submitted manuscript*, 2026. Studies stale rollouts, KL-direction sensitivity, teacher-cache constraints, estimator design, and throughput-quality trade-offs in asynchronous on-policy distillation pipelines.

Publications (* denotes equal contribution)

First-author / co-first-author publications

- **Kevin Galim***, Ethan Ewer*, Wonjun Kang, Minjae Lee, Hyung Il Koo, Kangwook Lee. Draft-based Approximate Inference for LLMs. *ICLR*, 2026.
- Wonjun Kang*, **Kevin Galim***, Seunghyuk Oh*, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, et al. ParallelBench: Understanding the Trade-offs of Parallel Decoding in Diffusion LLMs. *ICLR*, 2026.
- Wonjun Kang*, **Kevin Galim***, Yuchen Zeng*, Minjae Lee, Hyung Il Koo, Nam Ik Cho. State-offset Tuning: State-based Parameter-Efficient Fine-Tuning for State Space Models. *ACL*, 2025.
- **Kevin Galim***, Wonjun Kang*, Yuchen Zeng*, Hyung Il Koo, Kangwook Lee. Parameter-Efficient Fine-Tuning of State Space Models. *ICML*, 2025.
- Wonjun Kang*, **Kevin Galim***, Hyung Il Koo, Nam Ik Cho. Counting Guidance for High Fidelity Text-to-Image Synthesis. *WACV (Oral)*, 2025.
- Wonjun Kang*, **Kevin Galim***, Hyung Il Koo. Eta Inversion: Designing an Optimal Eta Function for Diffusion-based Real Image Editing. *ECCV*, 2024.

Other publications

- Seunghyuk Oh, Minjae Lee, **Kevin Galim**, Minseo Kim, Hyung Koo, Wonjun Kang, Hanbaek Lyu, Kangwook Lee. Inference-Aligned SFT for Diffusion LLMs via Group-based Trajectory Sampling. *ICLR DeLLa Workshop*, 2026.
- Minjae Lee*, Wonjun Kang*, Byeongkeun Ahn, Christian Classen, **Kevin Galim**, Seunghyuk Oh, Minghao Yan, Hyung Il Koo, Kangwook Lee. TABED: Test-Time Adaptive Ensemble Drafting for Robust Speculative Decoding in LVLMS. *EACL*, 2026.
- Wonjun Kang, Byeongkeun Ahn, Minjae Lee, **Kevin Galim**, Seunghyuk Oh, Hyung Il Koo, Nam Ik Cho. UNCAGE: Contrastive Attention Guidance for Masked Generative Transformers in Text-to-Image Generation. *IEEE Access*, 2026.
- Maxim Maximov, **Kevin Galim**, Laura Leal-Taixe. Focus on Defocus: Bridging the Synthetic to Real Domain Gap for Depth Estimation. *CVPR*, 2020.

Education

Technical University of Munich <i>Master's degree, Informatics: Games Engineering</i>	2016 – 2019 Grade: 1.4 (German system)
The University of Tokyo <i>Research, Computer Graphics</i>	2017 – 2018
Technical University of Munich <i>Bachelor's degree, Informatics: Games Engineering</i>	2013 – 2016 Grade: 1.6 (German system)

Skills

LLM Inference & Post-Training: Speculative Decoding · KV-Cache Optimization · Prompt/Context Compression · On-Policy Distillation · Efficient Inference

Machine Learning Research: Large Language Models · Diffusion Models · Parameter-Efficient Fine-Tuning · State Space Models · Diffusion LLMs

Computer Vision: Object Detection · Image Segmentation · Depth Estimation · Generative Models

Programming: Python · C++ · CUDA · PyTorch · TensorFlow · JavaScript

Languages

English (Full professional proficiency) · German (Native or bilingual proficiency) · Korean (Professional working proficiency)