

Ze-Wei (Johnny) Liou

✉ zwei.liou@princeton.edu 🌐 johnny-liou.github.io

RESEARCH INTERESTS

I work on efficient algorithms and systems for LLM inference, with a particular interest in linear attention (e.g., Mamba-2, GDN, KDA) and hybrid models.

EDUCATION

Sep 2025 – present	Princeton University Ph.D. in Electrical and Computer Engineering Advisor: Prof. Tri Dao	Princeton, NJ, USA
Sep 2020 – Jun 2024	National Taiwan University B.S. in Electrical Engineering Ranked 1st in three semesters (1st of 262 in the senior year)	Taipei, Taiwan GPA: 4.12 / 4.3

PUBLICATIONS

- 2026 **Ze-Wei Liou**, Harris Nover, Amey Naik, Andrii Skliar, Igor Shovkun, Ido Shahaf, and Tri Dao. “ReplaySSM: Cache SSM Inputs, Not State.” arXiv Preprint, 2026. [\[Paper\]](#) [\[NVIDIA Blog\]](#)
- Yung-Chin Chen, Chung Peng Lee, **Ze-Wei Liou**, and Naveen Verma. “Massive Spikes in LLMs are Bias Vectors: Mechanistic Uncovering and Spike-Free Quantization.” Findings of the Association for Computational Linguistics: EMNLP 2026. [\[Paper\]](#)
- 2025 **Ze-Wei Liou** and Ding-Yong Hong. “Optimizing Compute Core Assignment for Dynamic Batch Inference in AI Inference Accelerator.” Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC 2025). [\[Paper\]](#)
- 2023 G. W. Burr, H. Tsai, W. Simon, I. Boybat, S. Ambrogio, C.-E. Ho, **Ze-Wei Liou**, et al. “Design of Analog-AI Hardware Accelerators for Transformer-based Language Models (Invited).” IEEE International Electron Devices Meeting (IEDM 2023). [\[Paper\]](#)

RESEARCH ADOPTION

- 2026 **ReplaySSM**.
Faster decoding for SSM/linear attention layers in hybrid models by caching inputs instead of storing the state.
- Deployed in large-scale production models: **GLM-5.3-Flash** (Z.ai) and **Kimi-K3** (Moonshot AI).
 - Integrated into open-source ML frameworks such as **vLLM**, **TensorRT-LLM**, **FlashInfer**, and **SGLang**.
 - Featured on the [NVIDIA Technical Blog](#).

RESEARCH EXPERIENCE

Sep 2025 – present	Princeton University Graduate Research Assistant Advised by Prof. Tri Dao <i>Algorithm and system design for hybrid model decoding (ReplaySSM)</i> • Avoided per-step state writes for SSM/linear attention layers by representing the state as a checkpoint plus a buffer of recent inputs, halving state memory traffic during decoding. • Unlocked the separation of output computation from state updates: speculative decoding rollback becomes a pointer operation, and rounding error in a low-precision state no longer accumulates through the recurrence. • Achieved up to 1.47× and 2.1× speedups for standard and speculative decoding, respectively, over a tuned vLLM baseline at serving batch sizes, and up to 36× lower KL divergence with an NVFP4 state at 128K context.	Princeton, NJ, USA
Feb 2024 – Mar 2025	Academia Sinica Research Assistant Advised by Dr. Ding-Yong Hong <i>Efficient dynamic batch inference on multicore AI accelerators</i> • Proposed a dynamic-programming approach to compute-core assignment for dynamic batch inference, achiev-	Taipei, Taiwan

ing $3.05\times$ higher throughput than an EdgeTPU-style baseline.

Flexible hardware accelerator for N:M sparse models

- Designed a systolic-array FPGA accelerator supporting arbitrary N:M sparsity for both weights and activations.

Jun 2023 – Sep 2023

IBM Research, Almaden | Research Intern

San Jose, CA, USA

Advised by Dr. Geoffrey Burr and Dr. Hsinyu (Sidney) Tsai

Compute-in-memory (CIM) architectures for encoder-based Transformers

- Analyzed latency and area trade-offs between volatile and non-volatile memory-based CIM systems for BERT, and designed a ring-based dataflow enabling deep pipelining on volatile CIM that maximizes TOPS/s/mm².
- Validated data movement and numerical correctness in a C++ CIM simulator.

HONORS AND AWARDS

May 2023

Second Prize, National VLSI Circuit Design Contest, Taiwan

Spring 2023

Dean's List Award, National Taiwan University

Oct 2021

First Prize, National Meichu Hackathon, Taiwan