Hyunwoo Jae

hjae@umd.edu | +1 (227) 205-8858 | linkedin.com/in/hyunwooj2001 | medium.com/@hyunwooj2001

Education

University of Maryland, College Park Expected Fall 2026 – Expected Spring 2027
Master of Science in Computer Science

University of Maryland, College Park Spring 2024 – Expected Spring 2026
Bachelor of Science in Computer Science, Minor in Mathematics — Dean's List (2024 – 2025) GPA: 3.95/4.00

Research experience

Furong Huang Research Lab

Hold Onto That Thought: Assessing KV Cache Compression on Reasoning
Undergraduate Researcher May 2025 – Present

- Researching the impact of KV cache compression methods on LLM reasoning using benchmarks, while contributing to extending KVPress to support decoding-time compression.
- Designing systematic evaluations across Meta-Llama-3.1-8B-Instruct, Llama-3.1-Nemotron-Nano-8B-v1, and DeepSeek-R1-Distill models (Qwen-7B / Qwen-14B / Llama-8B), analyzing failure cases by examining correlations between average compression ratio, accuracy, and inference latency, and optimizing compression strategies.
- Implementing and optimizing the ShadowKV compression strategy through systematic hyperparameter tuning (compression ratio, eviction policy, max new tokens, and max context length), achieving reduced memory usage and latency while preserving reasoning accuracy across eight benchmarks (e.g., GSM8K, MATH500).
- Demonstrating that, for reasoning models, attention-based methods such as H2O and SnapKV-D achieve 80–121% of full-cache accuracy levels under 128–512 token cache budgets across eight reasoning benchmarks.

Tiamat-Benchmark: Embodied Multi-Modal Navigation and Question Answering
Undergraduate Researcher June 2025 - Present

- Designing and evaluating embodied agent tasks across seven scenarios using diverse 3D environments such as AI2-THOR, MP3D and Isaac-Sim 4.2.
- Leading outdoor scene generation in Isaac Sim 4.2 for large-scale autonomous robot navigation, developing automated pipelines to generate floor maps, simulate autonomous exploration, and produce synthetic navigation videos automatically paired with question–answer annotations using Python.
- Integrating MemoryEQA into the TIAMAT dataset by adapting its long-term exploration strategies for compatibility with outdoor scenes.
- Co-designing a question generation and validation pipeline spanning multiple reasoning levels — from object localization and attribute recognition to multi-step spatial planning and abstract goal inference using Python.

MORSE: A Suite of Programmatically Controllable Multimodal Reasoning Environments with Steerable Difficulty
Undergraduate Researcher Mar 2025 – July 2025

- Contributed to the MORSE framework and its MORSE-500 benchmark (ICLR 2025 submission), a video-centric suite of 500 curated problems across six reasoning dimensions (mathematical, spatial, temporal, physical, abstract, and planning), where state-of-the-art models achieved 23.6% accuracy compared to 55.4% human performance, revealing key reasoning deficits.
- Developed a physical reasoning dataset using Manim and Matplotlib to evaluate AI reasoning capabilities.
- Built a programmatically controllable difficulty-scaling pipeline parameterized by five orthogonal dimensions — entity complexity, reasoning depth, distractor density, temporal complexity, and visual complexity — enabling adaptive benchmark generation and curriculum design.
- Collaborated on dataset curation, ensuring high diversity and quality through human pilot testing.

UMD Bioinspired Robotics Lab

Hedwig: Flapping Wing Robot
Undergraduate Researcher Jan 2024 - Dec 2024

- Researched on developing Owl-Wing Drone ensuring safety through silent flight and fire detection.
- Developed the fire-detection pipeline by redesigning data preprocessing and model behavior to robustly handle noisy, low-resolution, and biased field video inputs.
- Collaborated with domain experts to define acceptable error tolerances and fine-tuned a YOLOv8 model to achieve 94% test accuracy under real deployment constraints.

Experience

Swell — *Researcher* May 2025 - Feb 2026

- Developing a Quantum Electronic Design Automation (Q-EDA) platform for superconducting circuits, unifying the fragmented quantum hardware design process.
- Responsible for designing full-stack architectures and developing AI tools that accelerate or complement traditional FEM simulation workflows.

Korean Undergraduate Student Association (KUSA) — *Event Chair | University of Maryland* Aug 2024 - Present

- Served as Event Chair, revitalizing a dormant community by re-registering KUSA and relaunching core events to support student adaptation to UMD.
- Organizing cross-major mentoring and study sessions to support students' academic, research, and internship goals.
- Expanded membership engagement by onboarding isolated students through online outreach and 1:1 invitations.

Econexenc — *Software Developer | Ori South Korea* Dec 2023 - Jan 2024

- Developed Document Management System in the PMS system to systematically manage strategy and operation by providing an overview of various projects, progress, manpower allocation, and document material at a glance utilizing Spring Boot framework, Thymeleaf template, Maria DB, SQL, HTML, CSS, and JavaScript.
- Designed a database schema and created a table that stores and allows for searching, reviewing, and accessing existing PDF manuals in the legal interpretation and technical reference library to provide above functionalities.

English for Life — *Software Developer | Seoul South Korea* Jan 2022 - Feb 2022

- Developed an English for Life web application that provides academy introduction, course enrollment, and inquiry system functionalities in Korean and English, utilizing HTML, CSS, JavaScript.

ROKA (Eunhye Community Childcare Center) — *Social Service Agent | South Korea* Mar 2022 - Dec 2023

- Served as Social Service Agent (alternative military) to 20 underprivileged children. The center provides child welfare services for the protection, education, and fostering in the community.
- Provided academic support (Math & English) and deployed fashion-style classifier as an engagement tool for the kids.

Publication

Minghui Liu, Aadi Palnitkar, Tahseen Rabbani, Kyle Rui Sang, **Hyunwoo Jae**, Dixi Yao, Shayan Shabihi, Fuheng Zhao, Tian Li, Ce Zhang, Furong Huang, and Kunpeng Zhang, "Hold Onto That Thought: Assessing KV Cache Compression on Reasoning," *NeurIPS 2025 Efficient Reasoning Workshop*, 2025. Submitted ICLR 2026

Zikui Cai, Andrew Wang, Anirudh Satheesh, Ankit Nakhawa, **Hyunwoo Jae**, Keenan Powell, Minghui Liu, Neel Jay, Sungbin Oh, Xiyao Wang, Yongyuan Liang, Tom Goldstein, Furong Huang, "MORSE: A Suite of Programmatically Controllable Multimodal Reasoning Environments with Steerable Difficulty", arXiv:2506.05523. Submitted ICLR 2026

Project

Dr. Teragotchi — *Group Project (Bitcamp 2024)* Apr 2025 - Apr 2025

- Developed a multi-platform mental health app leveraging AI for its core feature: virtual pets that perceive user emotions, interact adaptively, and evolve over time.
- Developed a FastAPI backend with Supabase and integrated OpenAI APIs to process emotional context, supporting personalized mental health journaling and interaction through carefully designed prompts and conversation logic.

Kissing Booth — *Group Project* Dec 2024 - Jan 2025

- Developed an AI-powered compatibility analysis app that allows users to assess their match potential with partners based on their photos, utilizing Tensorflow, Keras, Flutter, React-Native, AWS and SQL.
- Optimized a ResNet50-based deep learning model for photo similarity comparison, leveraging preprocessing and weighted loss adjustments and outlier detection on low-resolution, imbalanced data—achieving a loss reduction from 5.2 to 0.07.

Live Flight Tracker — *Personal Project* Feb 2023 - Mar 2023

- Developed a web application that displays real-time aircraft locations on a map and provides detailed information about flights through a popup and search function.
- Designed and developed a full-stack web application using React (frontend), Spring Boot (backend), and Python for data processing, integrating Oracle DB, Mapbox, and real-time flight data from the OpenSky API.

Technical Skills

Languages: Java, C, Python, Racket, x86-64 Assembly, HTML, CSS, JavaScript, SQL, SAS
Tools: PyTorch, TensorFlow, Keras, Isaac Sim 4.2, QT, Manim, Spring Boot, React, JMP Pro, Maria DB, Oracle DB, Pandas, NumPy, Matplotlib