HUI YEN, CHIA

+1(412) 579-0806 | huiyenc@andrew.cmu.edu | hychia88.github.io

Education

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Computational Design (Focus: Machine Learning Engineering) — *May 2026*

- **Relevant Coursework:** ML in Production, Principles of Software Construction, Data Structures & Algorithms, Intro to Deep Learning, WebApp Development.
- **Award:** CMU Architecture Merit Scholarship.

Tsinghua University — Beijing, China
Bachelor of Architecture — *June 2024*

Technical Skills

Languages: Python, Java, JavaScript, SQL, HTML/CSS.
Backend & Cloud: FastAPI, AsyncIO, Docker, Kubernetes, AWS (EC2, S3, RDS), Apache Kafka, PostgreSQL.
AI/ML Stack: PyTorch, FAISS (Vector DB), LangChain, RAG, Hugging Face, OpenAI API, Scikit-learn.
DevOps & Tools: CI/CD (GitHub Actions), PyTest, Prometheus, Grafana, Git, JWT Authentication.

Work Experience

Lennar Corporation — San Francisco, CA
Software Engineer Intern (AI/ML) — *June 2025 – August 2025*

- **High-Concurrency Microservices:** Architected and deployed scalable **FastAPI** microservices using **AsyncIO**; optimized throughput for multimodal payloads (image + text) by orchestrating concurrent OpenAI API calls, significantly reducing latency.
- **RAG Pipeline Engineering:** Engineered a Retrieval-Augmented Generation system using **LangChain** and a **FAISS** vector store; implemented hybrid retrieval (Dense + Sparse) to achieve **sub-second latency** on semantic search tasks.
- **Performance Optimization:** Developed an automated NLP pipeline incorporating few-shot classification and K-means clustering; processed and generated **20,000+ text embeddings in ¡300 seconds**, improving data ingestion speed by orders of magnitude.
- **System Reliability:** Enhanced system resilience against noisy production data by implementing exponential backoff strategies and strict Pydantic schema validation.

Projects

UniNest: Distributed Search Engine | *AWS, Docker, PostgreSQL* | GitHub Repo: UniNest — April 2025

- Engineered a distributed search backend combining **Postgres BM25** full-text search with semantic vector embeddings via **Reciprocal Rank Fusion (RRF)**, improving Precision@10 by 16% while maintaining **p95 latency at ∼329ms**.
- Deployed **JWT-secured microservices** to AWS EC2 via Docker; exposed RESTful endpoints tracking real-time metrics (p50/p99 latency) reaching ∼18 req/s throughput.
- Architected an automated ETL pipeline fetching ∼410 properties daily via scheduled jobs.

Real-Time Recommendation System | *Apache Kafka, MLOps, CI/CD* — October 2025

- Built a real-time user-event streaming architecture on **Apache Kafka** to enable closed-loop model evaluation.
- Productionized the model with a robust **CI/CD pipeline** including unit/integration tests and type checks; implemented **PSI-based Data Drift Monitoring** with automated alerting and rollback mechanisms.
- Designed a hybrid recommender (Collaborative Filtering + SVD) to mitigate cold-start issues, optimizing Recall@K metrics.

3T3D: Vision Transformer for 3D Generation | *PyTorch, ViT* | GitHub Repo: 3T3D — April 2025

- Implemented a **Vision Transformer (ViT)** encoder (DINOv2) and custom Transformer-based decoder in **PyTorch** to synthesize 3D-aware triplanar feature maps from 2D inputs.
- Optimized the 3D reconstruction module using the Marching Cubes algorithm for efficient mesh generation.