# 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: Applied / Production AI Systems)* *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*

**Award:** Tsinghua University Malaysian Outstanding Undergraduate Scholarship.

## Technical Skills

**Languages**: Python, Java, JavaScript, SQL, HTML/CSS.
**Backend & Cloud**: FastAPI, AsyncIO, Docker, AWS (EC2, S3, RDS), Apache Kafka (event streaming), 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:** Designed and implemented **FastAPI** microservices using **AsyncIO**; optimized throughput for multimodal payloads (image + text) by orchestrating concurrent OpenAI API calls with semaphore-based concurrency control, processing **1,000+ images in 240 seconds** via async batch processing.
- **RAG Pipeline Engineering:** Engineered a Retrieval-Augmented Generation system using **LangChain** and a **FAISS** vector store; implemented hybrid retrieval (Dense + Sparse + reranking) to achieve **sub-second latency** on semantic search tasks.
- **Performance Optimization and Reliability:** Built an automated NLP processing pipeline combining few-shot classification and clustering techniques, generating **20,000+ text embeddings in under 300 seconds**, improving data ingestion speed. 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 April 2025

- Built a hybrid search backend combining Postgres **BM25 full-text search** with semantic vector retrieval and **Reciprocal Rank Fusion (RRF)**, improving Precision@10 by 16% while maintaining **p95 latency ∼329ms**.
- Deployed **JWT-secured microservices** to AWS EC2 via Docker; exposed REST endpoints with real-time latency metrics; sustained ∼18 req/s throughput and automated ETL fetching ∼410 properties/day.

**Real-Time Recommendation System** | *Kafka, MLOps, Kubernetes (k3s)* October 2025

- Built a real-time user-event streaming pipeline on **Apache Kafka** with data quality gates (schema/type/range checks) and bad-record quarantine; monitored drift using **Population Stability Index (PSI)** on feature snapshots.
- Productionized retraining + deployment with **k3s CronJob** and **zero-downtime rollout** via atomic model hot-swap (versioned artifacts + symlink switch), enforced an online **latency eligibility gate** before promoting models.
- Built a hybrid recommender (CBF + UCF) with **RRF fusion** and **MMR re-ranking** for diversity; evaluated with leave-one-out Hit Rate plus service metrics (latency, artifact size).

**CAD-MLLM reproduction + Autocompletion** | *PyTorch, LoRA, Qwen-8B* | GitHub Repo November 2025

- Built data pipelines handling text, image, and point cloud inputs — alignment, dataloader design, and sampling for curriculum-based LoRA fine-tuning on a Qwen-8B backbone.
- Implemented CAD/mesh evaluation metrics covering sequence validity, geometry quality, topology, and enclosure; wrote reproducible inference scripts with rule-based guards for syntax and parameter sanity, verified through **OpenCascade/pythonocc** geometry builds.