HUI YEN, CHIA

https://hychia88.github.io/ +1(412) 579 0806 huiyenc@andrew.cmu.edu

EDUCATION

Carnegie Mellon University May 2026
Master of Science in Computational Design **Pittsburgh, PA, USA**
Courses: Machine Learning in Production, Principles of Software Construction, Data Structures and Algorithms for Applications, Intro to Deep Learning, WebApp Development

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

WORK EXPERIENCE

Applied AI Engineer Intern June -August 2025
Lennar Corporation **San Francisco, USA**

- Independently designed and deployed scalable FastAPI microservices that integrated construction domain requirements with multimodal AI workflows; enabled hybrid image search (semantic + keyword + reranking) across 1K+ images in ~240s through async batching and orchestration.
- Built high-performance vector database infrastructure using FAISS with O(1) metadata lookup and hybrid retrieval strategies, enabling real-time semantic search across construction domain knowledge with sub-second query response times.
- Engineered convo-to-report NLP pipeline processing construction conversations into structured data reports using few-shots classification, K-means/HDBSCAN clustering, and RAG framework with LangChain integration, achieving 20k+ text embeddings in under 300 seconds.
- Optimized ML system performance through asyncio implementation and concurrent OpenAI's API calls, delivering measurable precision/recall improvements on production-scale noisy text data while maintaining system reliability.

PROJECTS

UniNest AI-Powered Housing Search Engine - *Carnegie Mellon University* | github.com/uninest-ai/uninest April 2025

- Engineered hybrid search system combining Postgres BM25 full-text search with semantic vector embeddings (sentence-transformers) and Reciprocal Rank Fusion, improving Precision@10 from 0.16→0.42 (+16%) over keyword-only baseline while maintaining p95 latency ~329ms; exposed RESTful /metrics endpoint tracking p50/p95/p99 latency and ~18 req/s.
- Built multi-modal preference extraction pipeline leveraging Gemini API to parse user signals from chat conversations and uploaded images, generating structured property tags (validated with JSON schemas, exponential backoff retry logic) stored as versioned user preferences for downstream filtering
- Architected automated ETL pipeline fetching ~410 properties daily from Realtor APIs with scheduled jobs. Deployed JWT-secured microservices to AWS EC2 via Docker and storage via AWS S3.

3T3D 3D Generative AI Model - *Carnegie Mellon University* | github.com/1gfelton/3T3D.git April 2025

- Collaborated with team to research, design, and implement a novel 2D sketch-to-3D model generation pipeline using a Vision Transformer (ViT) architecture and triplanar representations for architectural applications.
- Contributed to the development of an encoder-decoder model leveraging a pre-trained DINOv2 (ViT) encoder and a custom Transformer/CNN-based decoder (PyTorch) to synthesize 3D-aware triplanar feature maps from input sketches.
- Developed the 3D reconstruction module to convert generated triplanar representations into surface meshes using the Marching Cubes algorithm (trimesh, Python).

Movie AI-Driven Recommendation System - *Carnegie Mellon University* October 2025

- Built real-time user-event streaming on Apache Kafka to enable closed-loop model evaluation and rapid iteration.
- Designed a hybrid recommender (CBF: TF-IDF + SVD + numeric features; UCF) fused via RRF and z-score blending; leave-one-out eval with Recall@K/NDCG: median rank 3 vs CBF 13; p90 27 vs UCF 97.
- Mitigated cold-start using demographic cohort priors and freshness-aware re-ranking, delivering >90% distinct lists for new users and fewer extreme misses.
- Productionized with automated data/model pipelines, CI/CD (unit & integration tests, linting, type checks), data-quality gates (schema/nullability/outliers), and PSI-based drift monitoring with alerting and rollback playbooks.

AWARDS & SCHOLARSHIP

Carnegie Mellon University Architecture Merit Scholarship 2024-2026

SKILLS

Languages & APIs: Python, Java, JavaScript
Databases & Cloud: PostgreSQL, MongoDB, AWS (EC2, S3, RDS)
AI/Retrieval: PyTorch, FAISS, RAG/LangChain, CLIP/ViT, NLP
DevOps & Quality: Docker, CI/CD, PyTest, Apache Kafka, OAuth 2.0/JWT
Framework: RESTful (FastAPI, Django, AsyncIO); Frontend (basic): React, Next.js, HTML/CSS
AEC / CAD: AutoCAD, Rhino/Grasshopper