Huang, Hejun

🌐 huanghejun.com | 🔗 hejunhuang | 📱 +1-616-274-8147 | ✉ huanghejun750@gmail.com | 🏠 Sunnyvale, CA

EDUCATION

M.Sc. in Aerospace, University of Michigan (Aerospace Department Fellowship)	08/2022 - 05/2024
M.Sc. in Mechanical, The Chinese University of Hong Kong	08/2019 - 12/2020
B.E. in Mechatronics, North China Electric Power University	08/2015 - 06/2019

WORK EXPERIENCE

Software Engineer at **Baidu USA** — 05/2024 - current

Multimodal Agent Collaborations for Video Generation [Demo] — Sunnyvale, California

- Pioneered a text-to-video pipeline integrating **LLM** with graphics APIs via **RAG** and **OpenAI API**.
- Executed **prompt engineering** to translate prompts into actionable steps for graphics engines using LLMs.
- Fine-tuned **Llama3** with RAG to generate rendering code from software documentation context.
- Generated videos from text prompts using **AI-generated code** in 3D modeling and animation software.
- Wrote testable application, keep **80%** code coverage with unit test via **Junit** and **Mockito** to ensure code quality.

Software Engineer at **Robotics and Autonomous Driving Lab, Baidu USA** — 01/2024 - 04/2024

Autonomous Cement Truck Development and Performance Optimization — Sunnyvale, California

- Developed planning, control, calibration modules on the Bazel-based **Apollo** platform using C++ and Python.
- Conducted CI/CD process of the AutoTruck for both open and closed-loop phase using **92GB** data in **4** months.
- Developed learning-based calibration tools, validated with GoogleTest, converting truck records into ProtoBuf format, accounting for load factors. This cut calibration times by **60%** and resulted in **2** patents.

Research Associate at **The Chinese University of Hong Kong** — 09/2020 - 06/2022

Online Educational Platform Development — Hong Kong SAR

- Designed and implemented Blackboard system using **GCP**, serving **100** faculty and TA in Engineering.
- Build a set of RESTful API for Assignment management models, serve assignment requests integrated with filtering, sorting and pagination.
- Deployed application with **Docker** integrated with **Kubernetes**, ensuring high availability and scalability.
- Boosted **authentication module** performance by **30%**, switched **JWT** token from **MySQL** to **Redis**.
- Built Frontend with **Thymeleaf** framework integrating with **HTML5**, **CSS3** and **JavaScript**.

PROJECT

Software Engineer at **FlashScale Shopping System** — 12/2022 - 12/2023

- Developed a **distributed** E-shopping system to manage flash sale event, handling traffic up to **10K QPS**.
- Optimized **peak load shifting** for the flash sale event using **Kafka**, handling **burst traffic** as **asynchronous** jobs to ensure system availability and reliability.
- Utilized distributed locking with **Redis** and **Lua** scripts to implement caching inventory lock and the **Try-Confirm-Cancel** pattern, effectively preventing overselling during sale events.
- Configured microservices with **Spring Cloud Gateway**, **OpenFeign**, **Consul** to enable service discovery/register, integrated with **AWS Load Balancer**/**Autoscaling** group on multiple **EC2 instance**, made service open to public with high availability.
- Improved search with **AWS OpenSearch**, reducing fuzz query latency by **30%** compared to MySQL.

SKILLSET

Development Languages: C++, Java, Python, SQL, MATLAB, Javascript, HTML5, CSS3

Software and Tools: Linux, Git, Docker, Kubernetes, Kafka, Spring Boot/MVC, MyBatis, Maven, React, GCP, AWS EC2/ASG/LB, AWS OpenSearch, MySQL, Spring Cloud Gateway/Consul, HuggingFace, Gradio, Bazel, Slurm, Protobuf, MongoDB, Jax, Tensorflow, PyTorch, TensorRT, LaTex, Blender, Autodesk Inventor.