


# Hejun (Kyven) Huang

🌐 [huanghejun.com](http://huanghejun.com) |  [hejunhuang](#) | 📞 +1-616-274-8147 | ✉ [huanghejun750@gmail.com](mailto:huanghejun750@gmail.com) | 🏠 Sunnyvale, CA

## EDUCATION

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

## WORK EXPERIENCE

<b>Software Engineer at Baidu USA</b>	05/2024 - current
Multimodal Agent Collaborations for Video Generation <a href="#">Demo</a>	Sunnyvale, California
<ul style="list-style-type: none"><li>Developed a text-to-video pipeline using <b>OpenAI API</b> and <b>Llama 8B</b>, optimizing for visual content via <b>RAG</b>.</li><li>Applied <b>prompt engineering</b> to convert text into video operations: Screenwriter, Scene Designer, Camera, Gaffer.</li><li>Managed <b>MongoDB</b> for 3D asset metadata, enabling dynamic scene composition with RAG-based LLM Director.</li><li>Integrated <b>Milvus</b> for vector-based code retrieval to enhance <b>Blender</b> Python API auto-coding for video generation.</li><li>Delivered this product for <b>ads</b> and <b>videos</b>, iteratively enhancing visual appeal based on <b>VLLM</b> agent feedback.</li></ul>	

<b>Software Engineer at Robotics and Autonomous Driving Lab, Baidu USA</b>	01/2024 - 04/2024
Autonomous Cement Truck Development and Performance Optimization	Sunnyvale, California
<ul style="list-style-type: none"><li>Developed calibration, planning and control modules on the <b>Bazel</b>-based <b>Apollo</b> platform using C++ and Python.</li><li>Conducted <b>CI</b> process of the AutoTruck for both open and closed-loop phase using <b>92GB</b> data in <b>4</b> months.</li><li>Delivered learning-based calibration tools, validated through simulation and real-car tests, reducing calibration times by 60% and securing 2 patents.</li></ul>	

<b>Research Associate at The Chinese University of Hong Kong</b>	09/2020 - 06/2022
Online Educational Platform Development <a href="#">Page</a>	Hong Kong SAR
<ul style="list-style-type: none"><li>Designed and implemented E-Learning system using <b>GCP/GKE</b>, serving <b>100</b> faculty and TA in Engineering.</li><li>Built <b>RESTful</b> APIs for Assignment Management, supporting integrated filtering, sorting, and pagination.</li><li>Deployed application with <b>Docker</b> integrated with <b>Kubernetes</b>, ensuring high availability and scalability.</li><li>Boosted <b>authentication module</b> performance by <b>30%</b>, switched <b>JWT</b> token from <b>MySQL</b> to <b>Redis</b>.</li><li>Built Frontend with <b>Thymeleaf</b> framework integrating with <b>HTML5</b>, <b>CSS3</b> and <b>JavaScript</b>.</li></ul>	

## PROJECT

<b>Software Engineer at FlashScale Shopping System</b>	12/2022 - 12/2023
<ul style="list-style-type: none"><li>Developed a <b>distributed</b> E-shopping system to manage flash sale event, handling traffic up to <b>10K QPS</b>.</li><li>Configured microservices with <b>Spring Cloud Gateway</b>, <b>OpenFeign</b>, <b>Consul</b> to enable service discovery/register, integrated with <b>AWS Load Balancer/Autoscaling</b> group on multiple <b>EC2 instance</b>, made service open to public with high availability.</li><li>Optimized <b>peak load shifting</b> for the flash sale event using <b>Kafka</b>, handling <b>burst traffic</b> as <b>asynchronous</b> jobs to ensure system availability and reliability.</li><li>Utilized distributed locking with <b>Redis</b> and <b>Lua</b> scripts to implement caching inventory lock and the <b>Try-Confirm-Cancel</b> pattern, effectively preventing overselling during sale events.</li><li>Improved search with <b>AWS OpenSearch</b>, reducing fuzz query latency by <b>30%</b> compared to <b>MySQL</b>.</li></ul>	

## 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, gRPC, MongoDB, Jax, Tensorflow, PyTorch, TensorRT, LaTeX, Blender, Autodesk Inventor.