CAI HONGYI

+60 1121029403 | xcloudfance@outlook.com | https://hongyicai.cc

Education

University of Malaya, Bachelor of Software Engineering Sept 2022 – July 2026

Skills

Programming Languages: Golang, Python, TypeScript, C#, C/C++, Lua, Dart, Java, Kotlin

Backend Frameworks: Django, Flask, Gin, FastAPI, Express, Springboot

Frontend: HTML/CSS/JS, jQuery, React.js, React Native, Flutter, Jetpack Compose (KMP), Astro, D3

Databases: PostgreSQL, MySQL, Redis, Firebase, MongoDB, ClickHouse

DevOps: Docker/Harbor, Git, Nginx, Linux, Kubernetes, Kubeflow, Kafka, Zookeeper, Prometheus & Grafana, Keepalived, Etcd, Keycloak (OAuth 2.0), Apache Spark

Cloud Platforms: Amazon AWS, Microsoft Azure, Alibaba Cloud, Tencent Cloud, Huawei Cloud (Modelarts, SWR, OBS)

Publications / Preprints

Pistachio: Towards Synthetic, Balanced, and Long-Form Video Anomaly Benchmarks
Li, J., **Cai, H.**, Dong, M., Pu, M., You, S., Wang, F., & Huang, T. (2025)
arXiv:2511.19474 (CVPR 2026 Under Review)
VLA-Pruner: Temporal-Aware Dual-Level Visual Token Pruning for Efficient Vision-Language-Action Inference
Liu, Z., Chen, Y., **Cai, H.**, Lin, T., Yang, S., Li, Z., & Zhao, B. (2025)
arXiv:2511.16449 (CVPR 2026 Under Review)
Evo-1: Lightweight Vision-Language-Action Model with Preserved Semantic Alignment
Lin, T., Zhong, Y., Du, Y., Zhang, J., Liu, J., Chen, Y., Gu, E., Liu, Z., **Cai, H.**,
Zou, Y., Zou, L., Zhou, Z., Li, G., & Zhao, B. (2025)
arXiv:2511.04555 (CVPR 2026 Under Review)
AutoDebias: An Automated Framework for Detecting and Mitigating Backdoor Biases in Text-to-Image Models
Cai, H., Rahman, M. M., Dong, M., Li, J., Pu, M., Fang, Z., Peng, Y., Luo, H., & Liu, Y. (2025)
arXiv:2508.00445 (CVPR 2026 Under Review)

Low-Confidence Gold: Refining Low-Confidence Samples for Efficient Instruction Tuning
Cai, H., Li, J.,Rahman, M.M., & Dong, W. (2025)
arXiv:2502.18978 (**EMNLP 2025 Findings**)

MergeIT: From Selection to Merging for Efficient Instruction Tuning
Cai, H., Fu, Y., Fu, H., & Zhao, B. (2025)
arXiv:2503.00034 (ACL 2026 Under Review)

AgileIR: Memory-Efficient Group Shifted Windows Attention for Agile Image Restoration
Cai, H., Rahman, M. M., Akhtar, M. S., Li, J., Wu, J., & Fang, Z. (2024)
arXiv:2409.06206 (**ICANN 2025 Proceedings**)

CFPFormer: Feature-pyramid like Transformer Decoder for Medical Image Segmentation
Cai, H., Rahman, M. M., Wu, J., & Deng, Y. (2024)
arXiv:2404.15451 (**IJCNN 2025 Proceedings**)

Work Experience

Technical Team Lead, Infinity Data Tech Sdn. Bhd. – Internship (Physical) Feb 2025 – Present

Java, Spring Boot, Spring Cloud, Kotlin, Jetpack Compose, Kubernetes, Cilium (eBPF), Kafka, Spark, ClickHouse, PostgreSQL, Redis, Prometheus/Grafana

- Led a 20-person full-stack team (backend, frontend, Android/iOS) to deliver multiple enterprise VPN and data products on time by establishing standardized CI/CD, code review, and cross-team Scrum processes
- Designed global VPN architecture with mTLS encryption, token-bucket rate limiting, and Keepalived high availability, eliminating session hijacking and DoS vulnerabilities for 500k+ MAU
- Implemented high-availability PostgreSQL cluster with read-write separation, streaming replication, and automated failover; integrated Flyway migrations and Redis caching, reducing P99 latency by 60% and enabling zero-downtime

schema evolution

- Architected and developed in-house distributed VPN node management system (Spring Boot + self-built control plane) supporting 1000+ global nodes: automated registration, configuration push, health checking, batch logging, keepalived orchestration, and one-click rolling upgrades
- Built full-funnel re-trackable analytics pipeline (Kafka → Spark Structured Streaming + Batch → ClickHouse), empowering product team with real-time conversion, retention, and LTV dashboards that directly drove multiple successful feature iterations, empowering user tag system and feature flag system.
- Planned and deployed production-grade bare-metal Kubernetes clusters with Cilium eBPF networking, Ingress-NGINX, Prometheus/Grafana; achieved 99.99% uptime and 40% lower infra cost than equivalent public cloud solutions

Research Assistant, Shanghai JiaoTong University – Internship (Physical) Sept 2024 – Sept 2025

Supervisor: Bo Zhao

- Trained multi-GPU fine-tuning pipeline for **LLaMA**, **Alpaca**, and **Vicuna** using **DeepSpeed**
- Spearheaded **data distillation mechanism** to spontaneously induce **instruction-tuning samples** from diverse contexts with high quality assessment
- Deployed and contributed to **ARM-based Ascend GPU training** on **Modelarts**, **Huawei Cloud**, involving containerized environment and prometheus supervision
- Led a **real2sim2real** project that reconstructs real objects using **GroundedSAM** + **Trellis** from video scanning into simulation environment, as well as attaching materials and attributes, enpowering **VLA models** with enriched data and synthetic 3D properties in MuJoCo emulator.
- Observed the visual-language retention strategy on Evo-1, without breaking the original understanding of visual grounding structure to gain efficient and effective architecture for Vision-Language-Action models
- Involved in VLA pre-training and downstream tasks fine-tuning, as well as tuning on emulation envrionemnt (e.g., **Copellia Sim, Issac Lab, MuJoCo** and etc.)

Research Assistant, TsingHua University – Internship (Remote) Apr 2024 – Sept 2024

Supervisor: Yan Wang

- Designed novel **multi-vehicle accident detection model** in complex **BEV scenarios** through innovative architecture.
- Conducted research on **activation quantization compression** and **model sparsity**, and led a project on **post-training activation-based weight compression** for **ViT**.
- achieving **4x model size reduction** while maintaining accuracy.
- Streamlined debugging workflow for **vision-language models**, improving development efficiency by 2x.

Machine Learning Engineer, 10 EPOCHS – Part-time (remote) Nov 2023 – Feb 2024

- Architected **SLURM cluster system** on existing GPU clusters optimizing **GPU resource allocation** for medicial imaging process.
- Addressed residual noise issues in high-resolution images through implementing **OpenCV-based pre/post-processing pipeline** for **medicial MRI images**, improving **noise reduction precision by 40%** with fine-grained control of **2dB denoising strength**.
- Designed **Canny Edge Detection** for optimizing **specific losses** during training medical image restoration for enhancing detail preservation.

Full-Stack Data Scientist, Overwatchs Technology – Internship (Physical) Mar 2023 – Sept 2023

NLP, Sentiment Analysis, Text Classification, Transformer, BERT, Transfer Learning, AWS SageMaker, MLOps, Docker, Kubernetes, Kubeflow, Computer Vision, Face Detection, Face Recognition, Anti-Spoofing, Liveness Detection

- Developed and deployed **NLP systems** using **Transformer**, **BERT**, and **XLNet**, achieving **85% accuracy** on **financial sentiment analysis**
- Leveraged **AWS SageMaker** for **MLOps deployment**, reducing model update cycles by 30% through automated CI/CD
- Employed finance-oriented analysis system through **Transfer Learning** and **Fine-tuning** on NLP and Vision models
- Architected **ML retraining system** using **Kubernetes** and **Kubeflow**, improving resource utilization by **40%**
- Built robust **KYC system** with **liveness detection** processing, and **Human Face Recognition** (**One-to-one**, **One-to-many verification**) with **ArcFace embeddings** in DynamoDB

Projects

Real2Sim2Real- A tool that migrates real-world objects in monocular video into 3D sim2real properties for VLA training. – Project Lead July 2025 – Sep 2025
github.com/MINT-SJTU/SIM2REAL2SIM

Python, MuJoco, Grounded SAM, Trellis, Pano2Room, OpenVLA

- Engineered a data pipeline utilizing Grounded SAM for accurate object segmentation and 3D reconstruction from monocular video, generating high-fidelity synthetic assets.
- Integrated the reconstructed assets into a physics-accurate sim2real environment built using MuJoco, enabling precise simulation for downstream tasks.
- Developed conversion modules (Trellis/Pano2Room methods) to attach rich 3D properties and attributes to assets, significantly expanding the scale and diversity of the dataset for OpenVLA training.

Verdant Search - Search Engine – Individual Project July 2020 – July 2021

github.com/xcloudfance/verdant_search

- Optimized **PostgreSQL search engine** handling concurrent access from distributed crawlers
- Engineered **Redis-based message queue** reducing indexing latency
- Implemented **full-text search** with specialized site filtering achieving **high QPS** under **high concurrency**