CAI HONGYI

+60 1121029403 | xcloudfance@gmail.com | hongyicai.cc | linkedin

Education

University of Malaya, Bachelor of Software Engineering — Sept 2022 – Feb 2027

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)
Machine Learning Fields: Generative AI / AIGC, Video / Vision-language Understanding, Efficient Natural Language Processing, Machine Unlearning, Data-centric AI, Medical Imaging, Low-level Vision

Publications / Preprints

Proceedings

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

When Vision Meets Texts in Listwise Reranking
Cai, H. (2026)
arXiv:2601.20623 **(SIGIR 2026 Long Paper Proceedings)**

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 Proceedings)**

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 Highlight)**

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)**

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)**

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)**

Under Review / Preprints

VisNec: Measuring and Leveraging Visual Necessity for Multimodal Instruction Tuning
Dong, M., **Cai, H.**, Li, J., Zhou, S., Ren, B., Peng, K., & Fu, Y. (2026)
arXiv:2603.01195 (ECCV 2026 Under Review)

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 (ECCV 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 (ICML 2026 Under Review)

Work Experience

Embodied AI Engineer, Devol Robots – Internship (Physical) Feb 2026 – Present

- Involved in V-jepa 2 latent world model style for Vision-Language Action models in robot arm controlling.
- Designed large MoE structure to align video vision, language prompt and action state in our own latent world model.
- Collected, trained and evaluated Pi-0.5 on Flexiv Horizon 4R robot arm and innovated on new control paradigm - force control, using force-torque data to precisely manipulating robot arm force and rotation.

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 team (backend, frontend, Android/iOS) to deliver multiple enterprise VPN and data products from 0 to 1 Million User, 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.
- 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 $\rightarrow$ Spark Structured Streaming + Batch $\rightarrow$ ClickHouse), empowering product team with real-time conversion, retention, and LTV dashboards that directly drove multiple successful feature iterations, for user tag system and feature flag system.
- Planned and deployed production-grade bare-metal Kubernetes clusters with Cilium eBPF networking, Ingress-NGINX, Prometheus/Grafana; Setting up P90, P99, and slow query warning systems.

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

Supervisor: Bo Zhao

- Designed and implemented data distillation framework that automatically induced 52k high-quality instruction-tuning samples from heterogeneous raw corpora, achieving new SOTA among data filtering methods on AlpacaEval, MMLU, GSM8K, etc.
- Pioneered and stabilized full-stack large model training on ARM-based Ascend 910 GPUs (8-card distributed training with DeepSpeed ZeRO-3, containerized environment, custom operator adaptation, and Prometheus+Grafana real-time monitoring on Huawei ModelArts
- Led Real2Sim2Real pipeline using Grounded-SAM and Trellis to reconstruct real-world objects from casual videos into simulation-ready 3D assets with accurate materials and physical attributes in MuJoCo, CoppeliaSim, IssacLab, significantly enriching simulated real-world training data for Vision-Language-Action models
- Investigated visual-language retention pre-training strategy on our own VLA model, Evo-1, explored maintaining original visual grounding performance

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

Supervisor: Yan Wang

- Designed end-to-end multi-vehicle accident detection model for complex BEV scenarios, achieving new SOTA on in-house large-scale dashcam dataset with a novel multi-scale perception architecture
- Led research and implementation of post-training activation-aware weight compression for Vision Transformers, successfully reducing model size by 4$\times$ with negligible accuracy drop
- Explored activation quantization and structured sparsity techniques; delivered a complete ViT compression toolkit that supports INT8/INT4 mixed precision and 50–70% sparsity without fine-tuning
- Optimized debugging and visualization pipeline for large vision-language models, cutting average iteration cycle time from 40min to under 20min and improving team development efficiency by 2$\times$

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

- Designed and deployed production-grade financial sentiment analysis system using BERT and XLNet; achieved 85% accuracy on in-house multi-language dataset and successfully served core risk-control business
- Built end-to-end MLOps pipeline with AWS SageMaker + automated CI/CD, reducing model iteration cycle from 2 weeks to 3 days
- Architected fully automated retraining platform based on Kubernetes and Kubeflow; supported dynamic scaling of NLP/CV models and improved cluster resource utilization by 40%
- Led development of real-time KYC system integrating liveness detection, 1:1/1:N face verification powered by ArcFace, and high-performance retrieval on DynamoDB; currently processing millions of verifications daily with 99.9% uptime

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

- Designed a multimodal listwise reranker extending classical **Learning-to-Rank (LTR)** paradigms to cross-modal evidence, jointly scoring text and visual content via **BLIP-2** embeddings over heterogeneous document collections
- Implemented a two-stage retrieval pipeline combining **BM25** and **Hierarchical Navigable Small World (HNSW)**-based **Approximate Nearest Neighbor (ANN)** search as a first-stage candidate pool, followed by a listwise reranker incorporating cross-modal vision-language features
- Evaluated against **MS MARCO** and **MMDocIR** benchmarks, demonstrating measurable ranking improvements over text-only **LambdaMART** and neural cross-encoder baselines by incorporating visual evidence into the reranking stage