Zhikang QIU

+86 18217533292 | +65 93971216 | zhikangqiu@gmail.com | LinkedIn | Singapore

EDUCATION

Shanghai Jiao Tong University | School of Electronic Information and Electrical Engineering, *Shanghai, China*
Master of Engineering GPA: 3.7 / 4.3 *Sep. 2018 – Mar. 2021*

Shanghai Jiao Tong University | School of Mechanical Engineering, *Shanghai, China*
Bachelor of Engineering GPA: 3.8 / 4.3 *Sep. 2014 – Jun. 2018*
Awards: Outstanding Graduate of Shanghai Jiao Tong University
Courses: Algorithm | Operation System | Computer Network | Machine Learning | Computer Vision | Data Mining
Areas of Expertise: LLM | Agent | Post-Training | Search Engine

WORK EXPERIENCE

TikTok, ***Senior Machine Learning Engineer*** *May. 2025 – present.*
- I work on Tako team, TikTok's in-app AI agent, where I build Tako's agent model, improve user experience, and mentor interns and new hires.
- **Honors/Awards:** 2025 Q3 Spot Bonus Award

Microsoft, ***Machine Learning Engineer 2*** *Nov. 2021 – May. 2025*
- Worked on the Bing Multimedia team at Microsoft AI, where I applied advanced AI technologies to enhance search quality and create innovative user experiences within the Bing search engine.

Baidu, ***Machine Learning Engineer*** *Apr. 2021 – Aug. 2021*
- Worked on the Video Search team, specializing in deep learning-based models to improve ranking quality.

SenseTime, ***Intern*** *Dec. 2019 – May. 2020*
- Contributed to the Autonomous Driving (AD) team by building vision models deployed within AD systems.

PROJECTS

Tako Global Expansion, *TikTok* *2025 – 2026*
- Goal: Support Tako's expansion into 9 countries across the Middle East, Southeast Asia, and Latin America.
- Owned agent model optimization for small-language markets. Produced small-language training data covering multiple target countries to improve experience in low-resource locales; reduced badcase rate from 19% to 15%. Independently completed model training, deployment, and crawling tests. Collaborated with cross-functional teams to ensure a successful launch.

Tako Agent Reply Style Optimization, *TikTok* *2025 – 2025*
- Goal: Optimize Tako's reply style in casual chat—shift away from an overly intimate, personified tone toward a warm, professional AI-assistant style.
- Compared Tako vs ChatGPT reply styles and designed style-related rubrics; in the data production pipeline, applied a critique-refine paradigm on these rubrics to optimize training data, then performed SFT and DPO on Tako agent's base model to adjust reply style. Overall experimental metrics remained flat; human evaluation in casual chat improved by +19/40 (ROW) and +18/40 (US).

Location-Aware Search Optimization, *TikTok* *2025 – 2025*
- Goal: Strengthen destination search within local services to better align with user intent and enhance user experience.
- Optimized destination recognition and retrieval in the search engine by conducting gap analysis and designing an end-to-end solution. Leveraged GPT and a 1.2B Thoth SLM, augmented with Google search results via RAG, to build destination recognition achieving 0.92/0.91 precision/recall in key countries, with the Thoth model nearly matching GPT-4o performance. Partnered with cross-functional teams to create an independent index and enhance recall and ranking, driving a 15% absolute lift in local search results and a 12% increase in video orders per search.

Video Moment Search, *Microsoft*, [*Demo*] *2023 – 2024*
- Goal: Enable users to instantly locate specific chapters or transcript sections within a video.
- Worked with other teams to build a large-scale chapter and transcript index using a custom chapter generation model, along with GPT and Whisper models. Designed and implemented a ranker to prioritize chapters and transcripts based on user queries. Implemented different ranking methods tailored to different query types. Resulting in a +1.2% improvement in Video CI and a +3.8% increase in PCR. Acquired extensive experience in taking a project from concept to release, cross-team collaboration as well as deep understanding of Bing search stack.

Video Chapter Generation Model, *Microsoft*, [*Demo*] *2023 – 2023*
- Goal: Develop a text generation model to split videos into timestamped chapters and generate title for each.
- Built an automated pipeline using Few-Shot Chain-of-Thought (CoT) prompting in GPT-4 to generate video chapters, each with precise start and end timestamps and a descriptive title. Finetuned multi-lingual LongT5 and LLaMA models with GPT-4-generated data, using video metadata and transcripts as inputs. Upgraded model for iterative generation on long videos. Employed GPT-4 to evaluate model performance based on Defect Rate (DR), achieving a DR of less than 7%. Deployed the model to GPU and CPU clusters, scaling throughput to over 10 million daily inferences. Gained extensive experience in LLMs.

Multi-modality Based Ranking Models for Bing Multimedia Search, *Microsoft* *2022 – 2023*

- Goal: Continuously improve the re-ranking model of Bing Multimedia Search for better results.
- The ranking model is a BERT based multi-modality model that leveraged query, doc and visual embedding as inputs. I built a new, more representative test set based on the latest Bing logs, improving evaluation precision. Fine-tuned teacher models with large scale parameters and implemented an ensemble of multiple teacher models to boost performance, then distilled it into a compact student model optimized for large-scale online inference training on billions of data points. Optimized the student model by TensorRT for large scale inference, enabling efficient online deployment. Achieved a +0.4% gain in Video DCG and +0.3% gain in Image DCG, enhancing overall search ranking performance.

Video Result Page Side-By-Side, *Microsoft* *2024 – 2025*

- Goal: Build a Side-By-Side metric to compare Bing Video and YouTube search result.
- Built an end-to-end configurable pipeline on Azure Data Factory to fetch and process Bing and YouTube search results, and automated the generation of evaluation tasks, reducing development cost by 90%. Maintained UHRS annotation quality to ensure fast and reliable metric production, and planned a next-generation evaluation system leveraging large language models.

Vehicles Key Points Detection and Direction Recognition, *SenseTime* *2019 – 2020*

- Developed a keypoint and direction detection model for tracking vehicles and estimating their intent. Built training datasets and designed a detection model using ResNet and FPN. Optimized the model by reducing channels to minimize latency while preserving performance. Serialized and deployed the compressed model in the AD system.

Video Action Recognition, *SJTU* *2019 – 2019*

- Proficient in commonly used deep learning algorithms for video action recognition, including TSN, I3D, and SlowFast, as well as related datasets. Trained a multi-modality model incorporating RGB frames, optical flow, and audio inputs, achieving top-5 accuracy of 86.6% on the Kinetics-600 dataset.

PUBLICATIONS

Zhikang Qiu, Xu Zhao, and Zhilan Hu. "Efficient temporal-spatial feature grouping for video action recognition." 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020.

SKILLS & OTHERS

Programming: Python, C++, C#, Go, JavaScript, Bash, PowerShell, SQL
AI Technologies: LLM, Agent, RAG, RLHF, PyTorch, TensorFlow, HuggingFace
Tools: Claude Code, OpenClaw, vLLM, DeepSpeed, Megatron-LM, ms-swift, Microsoft Azure, Azure ML, Grafana
Interests: Sports, Music, Electronics, Reading, Self-Driving Road Trips