# Zhaode Wang

**AI Inference Engine Expert · On-Device LLM · High-Performance Computing**

Email: hi@zhaode.wang · github.com/wangzhaode

## Education

| | | |
|---|---|---|
| **Institute of Computing Technology, Chinese Academy of Sciences**<br>State Key Laboratory of Computer Architecture | M.S. in Computer Architecture | 2017 - 2020 |
| **Shandong University** | B.S. in Computer Science and Technology | 2013 - 2017 |

## Experience

**Alibaba · Taotian Group** — **Senior Technical Expert · MNN Tech Lead** — 2020 - Present

- **MNN core architecture:** Own MNN's core architecture and technical roadmap across model conversion, graph optimization, memory planning, operator fusion, and CPU/GPU/NPU backends, delivering an end-to-end on-device inference engine spanning model import, compilation, execution, and heterogeneous acceleration; MNN has reached 15.8K+ GitHub stars.
- **MNN-LLM engineering:** Founded and currently lead MNN-LLM, building conversion and quantization tools, runtime, tokenizer, KV cache, evaluation, and ModelZoo; the stack supports 200+ language and multimodal models with over one million downloads. Delivered low-bit and mixed-precision quantization, weight/KV-cache mmap, FlashAttention, and speculative decoding for low-memory, long-context, and faster generation.
- **Production deployment:** Shipped on-device LLMs for Taobao feed, post-purchase recommendation, item reranking, and Xianyu message understanding at tens-of-millions-user scale with positive business impact; beyond LLMs, MNN's general on-device inference stack also supports Quark, Qwen, DingTalk, Youku, Taobao Instant Commerce, and Amap.
- **On-device SLM R&D:** Led a team to train a 0.3B SLM for on-device applications from scratch, spanning data collection, tokenization, pre-training, SFT, RL, evaluation, and MNN deployment.
- **On-device AI platform:** Addressed mobile deployment constraints for Python/OpenCV/NumPy-based algorithms by building an embedded Python runtime and model conversion, debugging, and release pipeline; enabled dozens of algorithms for Pailitao, Taobao Live, security, and feed, improving core CV/NumPy APIs by 15%+ on average while reducing a live-streaming app package by 13MB.
- **Engineering infrastructure:** Built MNN's documentation, cross-platform CI, accuracy/performance evaluation, and production crash analysis from scratch. Developed Agent/Skill workflows spanning coding, testing, performance analysis, and debugging, with model analysis, integration, export, and validation as a core workflow; reduced complex model integration from days of expert work to hours for engineers new to the stack.

**Cambricon Technologies** — **Compiler Engineer Intern** — 2018 - 2020

- Contributed to Cambricon's BANG compiler and linker, focusing on heterogeneous linking, and optimized Caffe operators for MLU accelerators.

**Megvii** — **Algorithm Engineer Intern** — 2019

- Contributed to traffic video understanding algorithms, focusing on lane marking detection, monocular-camera vehicle speed estimation, and traffic sign recognition.

**Microsoft Asia Engineering Academy** — **Software Engineer Intern** — 2016

- Evaluated LevelDB, Presto, and Hadoop for advertising data storage and analytics workloads.

## Selected Publications

| | | |
|---|---|---|
| Efficient LLM Inference on ARM via Hardware-Aware Operator Co-Design and Heuristic Mixed-Precision Search | ISCAS | 2026 |
| RecGPT-Mobile: On-Device Large Language Models for User Intent Understanding in Taobao Feed Recommendation | SIGIR | 2026 |
| MNN-LLM: A Generic Inference Engine for Fast Large Language Model Deployment on Mobile Devices | MM Asia | 2024 |
| Walle: An End-to-End, General-Purpose, and Large-Scale Production System for Device-Cloud Collaborative Machine Learning | OSDI | 2022 |
| LISA: Locality-Informed Speculative Decoding for Accelerating Autoregressive Image Generation | ECCV | 2026 |
| ARC-Decode: Accelerated Decoding with Risk-Bounded Acceptance | ICML | 2026 |
| Prism-MoE: Efficient Dense-to-MoE Conversion for Visual Autoregressive Generation | ICML | 2026 |
| AccKV: Towards Efficient Audio-Video LLMs Inference via Adaptive-Focusing and Cross-Calibration KV Cache Optimization | AAAI | 2026 |
| VertiKV: Vertical-Integrity KV Cache Compression for Efficient Multimodal Long-Context Inference | ACM MM | 2026 |
| MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference | ACL | 2025 |

## Patents & Applications

| | | |
|---|---|---|
| BANG-Linker: Heterogeneous Linking Method, Apparatus, and Related Products | Cambricon · Filed | 2019 |
| BANG Simulator: Programming Debugging Method, Apparatus, and Related Products | | |
| Solution for Deploying Large Language Models on Mobile Devices | Alibaba · Filed | 2024 |
| Saliency-Aware Blockwise Quantization via Floating-Point Zero-Point Degrees of Freedom | Alibaba · Pending | 2026 |
| Mixed-Precision Allocation for Low-Bit LLMs Based on Action-Conditioned Operator Error | | |

## Honors & Competitions

| | | |
|---|---|---|
| IEEE AICAS Grand Challenge: LLM Software and Hardware System Co-optimization | **First Prize** | 2024, 2025 |
| NPU Competitions: Tecorigin Operator Development & SOPHGO TPU Programming | **First Prize** | 2022, 2024 |

## Skills

| | |
|---|---|
| **Languages** | C, C++, Python, Assembly, CUDA, OpenCL, Vulkan, Metal, BANG, Java, Objective-C |
| **Expertise** | Compilers, AI inference engines, CPU/GPU/NPU operator optimization, model conversion, quantization, large language models |