Zebin Yang

13731109063 2401111912@stu.pku.edu.cn https://zebin-young.github.io/

ABOUT ME

I am a PH.D. student at SEC Lab at Peking University, advised by Prof. Meng Li. My current research interest lies in **Embodied AI**, **LLM Inference Acceleration**, **Model Compression**, and other emerging technologies. I received my bachelor's degree from Peking University under the supervision of Prof. Meng Li.

EDUCATION

Peking University, Beijing, China Sept. 2024 - Now

Ph.D., School of Integrated Circuits

Peking University, Beijing, China Sept. 2020 - Jun. 2024

B.S., School of Electronic Engineering and Computer Science **(with Honors)**

RESEARCH INTERNSHIPS & COLLABORATIONS

PrimeBot, Beijing, China Mar. 2026 - Jun. 2026

Advised by Jiafeng Xu, Hao Dong

Edge deployment and inference optimization for VLA models. Research output: **EagleVLA [1]**.

Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), Shenzhen, China Sept. 2024 - Sept. 2025

Advised by Bo Yu, Shaoshan Liu

Long horizon robot navigation and planning optimization. Research outputs: **EfficientNav [4]**, **KEEP [2]**.

RESEARCH INTERESTS

The listed references mean the relevant works I have participated in.

- **Efficient Embodied AI System**: Investigate efficient VLA inference and efficient embodied planning.
  - **Asynchronous VLA inference**: decouple VLA inference from action execution to enable continuous, real-time robot control, while addressing perception-execution misalignment and reducing reaction time through foresight-aligned correction and scheduling optimization [1].
  - **VLA inference engine**: develop EagleVLA-Edge, an efficient llama.cpp-based inference engine for real-time VLA deployment on edge platforms, including desktop GPUs, CPUs, NPUs, and NVIDIA Jetson Orin/Thor. On Jetson Orin, EagleVLA-Edge reduces $\pi_{0.5}$ latency by up to $3.44\times$ over PyTorch [1].
  - **VLA acceleration**: develop a layer-skipping framework that skips redundant computation according to action importance, achieving up to $3.75\times$ inference speedup and $85.7\times$ training-cost reduction while maintaining competitive manipulation performance [3].
  - **Memory management for embodied planning**: investigate efficient memory management methods for long-term embodied tasks such as navigation and household works, reducing planning latency by KV cache management, built on vLLM framework [2][4].
- **Model Compression & Inference Acceleration**: Design end-to-end optimization pipelines to improve inference efficiency across LLMs and linear transformers.
  - **KV cache management for LLM inference**: investigate efficient KV cache optimizations, including re-computation, position-independent caching, block attention [2][4].
  - **Quantization**: Optimize weight and activation quantization for fast inference; explore channel-adaptive precision allocation to balance model accuracy and computational overhead [5][6][8].
  - **Knowledge distillation (KD)**: employ logits-based distillation and feature-based distillation for model compression [5].
  - **Neural architecture search (NAS)**: automatically prunes redundant neurons for model compression [5].

PUBLICATIONS

** indicates equal contribution.*

The publication list is available through Google Scholar: please click here.

[1] **EagleVLA: Towards Onboard Real-Time Robot Control via Foresight-Aligned Asynchronous Inference.**
Zebin Yang, Qi Wang, Yunhe Wang, Xiurui Guo, Bo Yu, Shaoshan Liu, Jiafeng Xu, Hao Dong, Meng Li
CoRL'26 (Spotlight, Top 4.4%), GitHub 1, GitHub 2

[2] **KEEP: A KV-Cache-Centric Memory Management System for Efficient Embodied Planning.**
Zebin Yang, Tong Xie, Baotong Lu, Bo Yu, Shaoshan Liu, Meng Li
DAC'26, CCF-A, GitHub

[3] **DySL-VLA: Efficient Vision-Language-Action Model Inference via Dynamic-Static Layer-Skipping for Robot Manipulation.**
Zebin Yang, Yijiahao Qi, Tong Xie, Bo Yu, Shaoshan Liu, Meng Li
DAC'26, CCF-A, GitHub

[4] **EfficientNav: Towards On-Device Object-Goal Navigation with Navigation Map Caching and Retrieval.**
Zebin Yang, Sunjian Zheng, Tong Xie, Tianshi Xu, Bo Yu, Fan Wang, Jie Tang, Shaoshan Liu, Meng Li
NeurIPS'25, CCF-A, GitHub

[5] **MCUBERT: Memory-Efficient BERT Inference on Commodity Microcontrollers.**
Zebin Yang*, Renze Chen*, Taiqiang Wu, Ngai Wong, Yun (Eric) Liang, Runsheng Wang, Ru Huang, Meng Li
ICCAD'24, CCF-B

[6] **FastQuery: Communication-efficient Embedding Table Query for Private LLM Inference.**
Chenqi Lin*, Tianshi Xu*, **Zebin Yang***, Runsheng Wang, Ru Huang, Meng Li
DAC'24, CCF-A

[7] **ProPD: Dynamic Token Tree Pruning and Generation for LLM Parallel Decoding.**
Shuzhang Zhong, **Zebin Yang**, Ruihao Gong, Runsheng Wang, Ru Huang, Meng Li
ICCAD'24, CCF-B

[8] **LightMamba: Efficient Mamba Acceleration on FPGA with Quantization and Hardware Co-design.**
Renjie Wei*, Linfeng Zhong*, Songqiang Xu*, **Zebin Yang**, Qingyu Guo, Yuan Wang, Runsheng Wang, Meng Li
DATE'25, CCF-B, GitHub

HONORS AND ACHIEVEMENTS

National Scholarship (Ph.D Students) — Sep. 2026
Outstanding Graduates at Peking University — Jul. 2024
Merit Student — Sep. 2022,2023
UBIQUANT scholarship — Dec. 2023
Lee Wai Wing scholarship — Dec. 2022

COMPETITION AWARDS

The 13th Chinese Mathematics Competitions (CMC) — 2022
First Prize
2022 Contemporary Undergraduate Mathematical Contest in Modeling (CUMCM) (Higher Education Press Cup) — 2022
Second Prize
The 38th National Regions Physics Competition — 2022
Second Prize
2021 Asia and Pacific Mathematical Contest in Modeling (APMCM) — 2021
Second Prize