



Hangliang Ding

✉ pianoqwz@gmail.com ◇  github.com/foreverpiano ◇  foreverpiano.github.io

EDUCATION

Bachelor of Science in Mechanics

Sept. 2021 - present

Xingjian College, Tsinghua University, GPA: 3.66 / 4.00

Relevant Courses:

Operating Systems (93), High Performance Computing and Parallel (94), Object-Oriented Programming (98)
Ordinary Differential Equations (100), Tensor Analysis and Differential Geometry (94)

RESEARCH EXPERIENCE

Efficient-vDiT: Efficient Video Diffusion Transformers With Attention Tile

June 2024 - present

Supervised by Prof. Hao Zhang University of California San Diego

- Developed a novel pipeline for accelerating a video diffusion model that with only 1% extra pretraining FLOP, achieves a **7.8**× inference speedup on a single GPU with minimal video degradation.
- Adopted consistency distillation and sparse attention distillation, with a specialized fast sparse attention kernel and context parallel methods, enabling a further **30**× inference speedup on 4 GPUs.

MiniKV: 2-Bit Layer-Discriminative KV Cache Compression

Feb. 2024 - June 2024

Supervised by Prof. Minjia Zhang University of Illinois Urbana-Champaign

- Pioneered a novel framework that significantly reduces computational memory and latency in LLMs with **86%** KV cache compression ratio while retaining 98.5% accuracy for long-context tasks.
- Optimized the KV cache eviction by combining a 2-bit quantized mechanism with a specialized CUDA selective flash-attention kernel and adjusting cache size according to adaptive layer-wise strategies.

AutoGLM & AgentBench: Web Navigating Agent and Evaluation Framework

Mar. 2023 - Jan. 2024

Intern at ZhipuAI / ChatGLM

- Classified real-world browsing options and designed auto-collected browsing traces data framework, building a more efficient language model-driven automated web navigation agent.
- Designed comprehensive text-based interactive environments for real-world cases and built a distributed evaluation framework to assess the performance of LLMs in agent tasks.

PUBLICATION

Efficient-vDiT: Efficient Video Diffusion Transformers With Attention Tile **In submission**

Hangliang Ding, Dacheng Li, Runlong Su, Zhijie Deng, Ion Stoica, Hao Zhang

MiniKV: Accelerating LLM Inference via 2-Bit Layer-Discriminative KV Cache **In submission**, arxiv 2411.18006

Akshat Sharma, **Hangliang Ding**, Jianping Li, Daniel Neel, Minjia Zhang

AgentBench: Evaluating LLMs as Agents **Accepted in ICLR 2024**, Github ★ 2300+

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, **Hangliang Ding**, Kaiwen Men et al.

AWARDS & HONORS

Gold Medal of ICPC (Algorithm Programming Contest) Xi'an regional

Nov. 2022

Gold Medal of ICPC (Algorithm Programming Contest) Nanjing regional

Nov. 2021

Tsinghua Scientific Research Excellence Scholarship

Dec. 2023

Tsinghua Scientific Research Excellence Scholarship

Dec. 2022

SKILLS

Computing skills: CUDA, OpenAI Triton, C/C++, Python, Golang

Frameworks: Vllm, Pytorch, DeepSpeed, CUDA Graph, Accelerator, Huggingface Transformers