Shuo Yang

andy_yang@berkeley.edu https://andy-yang-1.github.io/

Education

Shanghai Jiao Tong University — Shanghai, China
Bachelor of Computer Science — *Sept. 2020 - June. 2024*

- Member of ACM Honors Class, which is an elite CS program for top 5% talented students
- GPA (All): 91.53/100, **Ranking: 2/36**

University of California, Berkeley — Berkeley, CA
EECS PhD advised by Prof. Ion Stoica — *Sept. 2024 - Present*

- Member of Berkeley Sky Computing Lab (RiseLab)
- Member of LMSYS

Research

Systems for large language models

- Efficient online & offline serving: BlendServe (ASPLOS26), S-LoRA (MLSys24), Prism, Concerto and UCCL
- System-algorithm co-design: DoubleSparsity, HashAttention (ICML25), VAttention (ICLR26) and Twilight (NeurIPS25 Spotlight)

Systems and algorithms for video generation and world models

- Efficient and high-quality video generation: Sparse VideoGen (ICML25), Sparse VideoGen2 (NeurIPS25 Spotlight) and SLA (ICLR26)
- Long-horizon video generation and world models: Quant-VideoGen, StreamDiffusionV2 (MLSys26) and RadialAttention (NeurIPS25)
- Open benchmarks and datasets: Worldmodelbench (NeurIPS25)

Experience

Amazon — Santa Clara, CA, USA
Research Scientist Intern, advised by Yida Wang — *May. 2025 - Aug. 2025*
Research Topic: Efficient high-quality long video generation

University of California, Berkeley — Berkeley, CA, USA
Research Assistant, advised by Prof. Ion Stoica — *Mar. 2023 - Dec. 2023*
Research Topic: Low latency inference for LLM; Fast LoRA serving system; Contamination study of LLM

Shanghai Jiao Tong University — Shanghai, China
Undergraduate Researcher, advised by Prof. Tianqi Chen — *July. 2022 - Mar. 2023*
Research Topic: Building fast and efficient operators and providing strategies to generate the operators

Selected Publications

BlendServe: Optimizing Offline Inference for Auto-regressive Large Models with Resource-aware Batching
Y. Zhao, **S. Yang***, K. Zhu, L. Zheng, B. Kasikci, Y. Zhou, J. Xing, I. Stoica(*equal contributions)*

- Accepted by *ASPLOS 2026* (all 3 badges in Artifact Evaluation)
- BlendServe exploits the relaxed latency requirements in offline batch inference to reorder and overlap requests with varied resource demands while ensuring high prefix sharing.

Sparse VideoGen: Accelerating Video Diffusion Transformers with Spatial-Temporal Sparsity
H. Xi, **S. Yang***, Y. Zhao, C. Xu, M. Li, X. Li, I. Stoica, K. Keutzer, S. Han(*equal contributions)*

- Accepted by *ICML 2025*

- SVG reveal that the attention heads can be classified into spatial head and temporal head. SVG proposes an online profiling strategy to predicts the type of attention head. Combined with layout transformation, SVG achieves up to 2.33x end-to-end speedup on HunyuanVideo while preserving generation quality

Sparse VideoGen2: Accelerating Video Diffusion Transformers via Semantic-Aware Permutation

S. Yang*, *H. Xi**, Y. Zhao, C. Xu, M. Li, X. Li, K. Keutzer, S. Han, I. Stoica(*equal contributions)*

- Accepted by *NeurIPS 2025* as **Spotlight**
- SVG2 identifies two failure reasons of existing BSA: (1) Inaccurate identification (2) Computation waste. To bridge this gap, SVG2 proposes semantic-aware permutation to cluster and reorder critical tokens into a contiguous layout. Combined with flash-kmeans, SVG2 achieves a new pareto frontier in generation quality and efficiency

Quant VideoGen: Auto-Regressive Long Video Generation via 2-Bit KV-Cache Quantization

*H. Xi**, ***S. Yang****, *Y. Zhao, C. Xu, M. Li, X. Li, I. Stoica, S. Han, K. Keutzer(*equal contributions)*

- QVG identifies that KV-cache is bottlenecking auto-regressive video generation. To bridge this gap, QVG exploits video's spatiotemporal redundancy via Semantic-Aware Smoothing to produce low-magnitude, quantization-friendly residuals. QVG reduces KV memory by 7.0× with significantly better quality over baselines

Flash-KMeans: Fast and Memory-Efficient Exact K-Means

S. Yang*, *H. Xi**, Y. Zhao, M. Li, K. Keutzer, S. Han, C. Xu, I. Stoica(*equal contributions)*

- Flash-KMeans revisit k-means algorithm under the lens of modern AI system design and enable k-means as an online primitive. It point out that existing GPU implementations of k-means remain fundamentally bottlenecked by low-level system constraints rather than theoretical algorithmic complexity. Evaluations demonstrate that flash-kmeans achieves up to 17.9× end-to-end speedup over best baselines, while outperforming industry-standard libraries like cuML and FAISS by 33× and over 200×, respectively.

Projects

Sparse VideoGen

Training-free acceleration framework for high-quality video generation.

Code for the **Sparse VideoGen** and **Sparse VideoGen2**, 600+ stars on github

Flash K-Means

Fast and memory-efficient exact K-Means

Fast batched K-Means clustering implemented with Triton GPU kernels, 16× faster than prior SOTA

LLM Decontaminator

Decontamination system for language model datasets. (LMSYS's 4th Project)

Code for the paper "Rethinking Benchmark and Contamination for Language Models with Rephrased Samples", adopted by ReflectionAI

Fellowship & Scholarship

Fellowship

- 2025 Amazon AI PhD Fellowship
- 2024 Berkeley EECS PhD Fellowship

Scholarship

- 2022 Hanying-Juhua Scholarship
- 2021 Ruiyuan-Hongshan Scholarship
- 2020, 2021, 2022 Zhiyuan Honorary Scholarship

Other Experience

Great Minds in Computer Science CS1950 — Shanghai, China
Teaching Assistant — *June. 2021 - Dec. 2021*

Probability Theory MATH2701 — Shanghai, China
Teaching Assistant — *June. 2022 - Dec. 2022*