

Jinyuan Mao

jymao@zju.edu.cn | [homepage](#)

EDUCATION

Zhejiang University

B.S. in Statistics & B.Mgt. in Info. Management and Info. Systems (Dual Degree)

Hangzhou, China

Sep. 2023 – Jun. 2027

- **GPA:** 92.5/100, **Rank:** 1/20
- **Research Interests:** Generative Models, Efficient AI, Computer Vision

PUBLICATIONS

* denotes equal contribution

[1] **BIFE: Better Interaction, Fewer Errors for Minute-Long Video Generation**

Zeyu Zhang*, **Jinyuan Mao***, Shuning Chang, Yuanyu He, Yizeng Han, Jiasheng Tang, Fan Wang, Bohan Zhuang
SIGGRAPH Asia 2026 Technical Communications | [arXiv](#)

[2] **WorldAttention: An Efficient Attention Architecture for Interactive Video World Models**

Zeyu Zhang*, **Jinyuan Mao***, Dakai An*, Wangbo Zhao, Hanfeng Lu, Jiasheng Tang, Yinghao Yu, Wei Wang, Bohan Zhuang
NeurIPS 2026 Reviewed (5/5/5 – Accept/Accept/Accept) | [arXiv](#)

[3] **HimmBO: High-Dimensional Interactive Model-Merging Bayesian Optimization for Image Generation**

Jinyuan Mao, Zhihao Shen, Chenxi Liu
Under review at Eurographics 2027

RESEARCH EXPERIENCE

Research Intern, PixAI

Interactive image generation; advised by Chenxi Liu (University of Victoria)

Jun. 2026 – Oct. 2026

Hangzhou, China

- Led **HimmBO** [3] as first author, owning the method design, implementation, experiments, and writing.
- Proposed a two-stage preferential Bayesian optimization framework that scales interactive adapter (LoRA) merging from a few dozen to **1080 adapters**, with a level-set screening stage that needs no image generation during interaction.
- Designed a style-aware kernel that transfers preference information across visually similar adapters; reached a given similarity with **under 1/10** of the image generations required by prior work.

Research Intern, ZIP Lab, Zhejiang University

Efficient video generation and world models; advised by Prof. Bohan Zhuang

Oct. 2025 – Jun. 2026

Hangzhou, China

- **WorldAttention** [2]: observed highly heterogeneous sparsity patterns across attention heads in video diffusion transformers, and designed **Hybrid Sparse Attention**, pairing a linear global branch with head-adaptive block-sparse attention. Implemented the framework and system-level evaluation, achieving a **2.21×** end-to-end inference speedup while maintaining quality.
- **BIFE** [1]: co-designed the semantic sparse KV cache for retrieval-based long-range conditioning and ran much of the experimental validation; improved VDE-Subject by **22.2%**. Led the rebuttal, drafting per-reviewer responses and running supplementary experiments.

Project Leader, Provincial SRTP Project

Database Agent Framework

Apr. 2024 – Jun. 2025

Zhejiang University

- Proposed a “Two-Stage Trial-and-Error Learning” framework to improve database agents’ interaction logic, yielding clear gains on research benchmarks; filed one national invention patent.

HONORS & AWARDS

National Scholarship (Top 0.2%), three times

2024, 2025, 2026

1st Place Globally, SC25 IndySCC Student Cluster Competition

Nov. 2025

National 2nd Prize, 12th National Parallel Application Challenge (PAC)

Aug. 2025

2nd Prize, ASC25 Student Supercomputer Challenge

May 2025

1st Place, National Finals, ICT Competition

Mar. 2025

OTHER EXPERIENCE

AI Engineering Intern, Nodesk AI

Jun. 2025 – Aug. 2025

Industry internship

Hangzhou, China

- Built a smart-reply cross-validation module, improving accuracy by **15%** via fine-tuning; co-authored a patent on a guardrail mechanism combining reranking models with scenario recognition.

Operations & Infrastructure Lead, ZJU X-Lab

Jun. 2025 – Present

Student-run innovation organization

Zhejiang University

- Manage the GPU clusters with Docker-isolated environments for 10+ projects; built a private Overleaf, NAS storage, VPN, and the organization's website.

SKILLS

Languages: Python, C/C++, SQL, $\text{\LaTeX}$

ML & Systems: PyTorch, Diffusers, MPI, OpenMP, Docker, Linux