Benhao Huang

huskydogewoof@gmail.com

https://huskydoge.github.io/ | Google Scholar

EDUCATION

School of Electronics Information and Electrical Engineering, Shanghai Jiao Tong University Sep 2021 – Present

B.ENG. in Computer Science and Technology (IEEE Honor Class), GPA: 93.14/100.00, 4.08/4.30, (Rank 5/127)

Relevant Coursework:

- **Computer Science:** Design and Analysis of Algorithms (A+), Computer Networks (A+), Operating System (A+), Programming Languages and Compilers (A+), Natural Language Processing (A+), Database System Technology (A+), Computer Vision (A+), Principles and Methods of Program Design (A+), Program Design Practice (A+), Introduction to Data Science (A+)
- **Mathematics:** Mathematical Analysis (A+), Linear and Convex Optimization (A+), Information Theory (A+), Complex Analysis (A+), Probability and Statistics (A+)
- **Additional Academic Pursuits:** Engaged in a dual degree program in Mathematics and Applied Mathematics. Coursework included: Complex Analysis, Abstract Algebra, Linear Algebra II

TOEFL: 107, S24, R27, L29, W27, **GRE**: V157, Q170, AW4.0

RESEARCH INTERESTS

- Multi-modal World Model, World Model-based Reasoning and Planning, VLM / LLM Agents
- Interpretability AI, Data Influence Analysis, Dataset Synthesis and Distillation
- Acceleration of Generative Model

PUBLICATIONS

- DCA-Bench: A Benchmark for Dataset Curation Agents [paper]
  Benhao Huang, Yingzhuo Yu, Jin Huang, Xingjian Zhang, Jiaqi W. Ma. (Under review at ICLR 2025)
- Contrasting Adversarial Perturbations: The Space of Harmless Perturbations [paper]
  Lu Chen, Shaofeng Li, ***Benhao Huang***, Fan Yang, Zheng Li, Jie Li, Yuan Luo. **AAAI 2025**
- Defining and Extracting Generalizable Interaction Primitives from DNNs. [paper | code]
  Lu Chen, Siyu Lou, ***Benhao Huang***, Quanshi Zhang. **ICLR 2024.**

SELECTED PROJECTS

PandoraV2: Towards General World Model with Natural Language Actions and Video States

In Progress (co-lead) *Advisor: Prof. Zhiting Hu* [code] Jun 2024 – Present

- Diffusion Game Engine: Built an auto-regressive Image-to-Video (I2V) model capable of simulating 2D platformer games (e.g., Mario), allowing control of both characters and environmental elements using text inputs on the fly. Proposed and implemented window-slide conditioning to support the generation of game videos lasting longer than one minute.
- Video Diffusion Model Acceleration: Spearheaded a sub-project focusing on optimizing video diffusion for real-time game generation, achieving generation speeds of under 1 second per round.
- Complex Video Captioning: Led a sub-project aimed at enhancing video captioning for complex scenarios where even state-of-the-art visual language models tend to falter, ensuring more accurate descriptions.
- Large-Scale Training Data Pipeline: Designed and implemented a high-efficiency processing pipeline for video training data, processing over 10 million videos simultaneously, significantly improving the overall data quality and processing speed.

DCA-Bench: A Benchmark for Dataset Curation Agents

ICLR 2025 Under Review (1st author) *Advisor: Prof. Jiaqi Ma* [paper] Jan 2023 – Present

- Identified a novel task for LLM agents – detecting dataset quality issues (e.g. cross-file discrepancy, hidden corruptions) for the purpose of automating AI training data curation – and developed the first benchmark for this task. We collected 221 instances and designed a four-level difficulty for more fine-grained analysis on agents' performance.
- Developed a LLM-based automatic evaluator for scalable evaluations, achieving a reliable and robust to self-preference or length bias demonstrated through comprehensive experiments.
- Led the project, conducting surveys, implementing code, completing experiments, and writings.

Defining and Extracting Generalizable Interaction Primitives from DNNs

ICLR 2024 *Advisor: Prof. Quanshi Zhang* [paper | code] Sep 2023 – Jan 2024

- Ideated a method to extract shared interactions of different DNNs trained for the same task.

- Conducted contrast experiments, illustrating that the extracted interactions can better reflect common knowledge shared by different DNNs. These shared interactions can be further used to interpret the output from DNNs.
- Implemented the main experiment codes and engaged in algorithm design, deploying the GitHub repository.

RESEARCH EXPERIENCE

- Research Intern, MAITRIX Lab, University of California San Diego. Apr 2024 – Present

Developed visual World Model with focus on game generation *Advisor: Prof.* *Zhiting Hu*

- Research Intern, Alignment Team, Moonshot AI Mar 2024 – Jun 2024

Explored prompt priorities alignment of LLMs *Advisor:* *Flood Sung*, *Yanan Zheng*

- Research Intern, TRAIS Lab, University of Illinois Urbana-Champaign Nov 2023 – Present

Constructed a LLM Agent benchmark for dataset issue detection *Advisor: Prof.* *Jiaqi Ma*

- Research Intern, XAI Lab, Shanghai Jiao Tong University Apr 2023 – Jan 2024

Extracted common knowledge of different LLMs *Advisor: Prof.* *Quanshi Zhang*

AWARDS

• National Scholarship (Top 0.2% nationwide, ranked 1/127 in my major)	2023 - 2024
• Rui Yuan-Hong Shan Scholarship (Top 2%), SJTU	2022 - 2023
• Shao Qiu Scholarship (Top 4%), SJTU	2021 - 2022
• Meritorious Winner of Mathematical Contest in Modeling	2022

OTHER

• Student Mentor for CS2612 Programming Languages and Compilers	2023 - 2024
• Student Mentor for CS2601 Convex Optimization	2023 - 2024
• Volunteer at Shanghai Marathon	2022 - 2024
• Member of the Outreach Department, SJTU Spark Program Student Association	2021 - 2022