WENTAO HU

Phone: (+86) 18980399950 ◇ Email: wentao002@e.ntu.edu.sg
Homepage: wentaohu1208.github.io
Github ◇ LinkedIn

EDUCATION

Nanyang Technological University (QS15) *Jul. 2023 – Nov. 2025 (expected)*
M.Eng in Computer Science and Engineering
GPA: 3.83/5.0 Supervisor: Prof.Hanwang Zhang

Hunan University (Project 985) *Sep. 2019 – June. 2023*
B.Sc in Statistics
GPA: 3.55/4.0

RESEARCH INTERESTS

Image Generation, Multimodal Large Language Model, Video Generation

PUBLICATIONS

[1] W. Lin*, L. Jia*, **W. Hu***, K. Pan, Z. Yue, W. Zhao, J. Chen, F. Wu, H. Zhang. *Reasoning Physical Video Generation with Diffusion Timestep Tokens via Reinforcement Learning.* arXiv:2504.15932, Under Review, 2025. (*Co-first authors)

[2] B. Wang, Z. Yue, F. Zhang, S. Chen, L. Bi, J. Zhang, K. Chan, J. Pan, W. Wu, M. Zhou, W. Lin, K. Pan, S. Zhang, L. Jia, **W. Hu**, W. Zhao, H. Zhang. *Selftok: Discrete Visual Tokens of Autoregression, by Diffusion, and for Reasoning.* arXiv:2505.07538, Under Review, 2025.

[3] **W. Hu**, et al. *On Path to Multimodal Generalist: Levels and Benchmarks.* arXiv:2505.04620, *International Conference on Machine Learning (ICML)*, 2025.

RESEARCH EXPERIENCE

Huawei Central Media Technology Institute, 2012 Laboratory Aug. 2024 - Present
Research Intern Singapore

Autoregressive Video Generation with Reinforcement Learning Aug. 2024 - Mar. 2025

- **Background**: Existing video generation models struggle to adhere to physical laws.
- **Method**: We analyze the limitations of current spatial tokenizers in autoregressive models; train an autoregressive model with a DDT tokenizer; And apply GRPO for reinforcement learning-based fine-tuning, enabling symbolic reasoning in the visual domain and thus achieving physical extrapolation.
- **Results**: We achieve near-perfect out-of-distribution prediction across three basic motion types (uniform motion, parabola, and collision).
- **My Involvement**: Conducted an extensive literature review on video generation model. Participated in the experimental design. Based on the phyworld dataset, curated and constructed a custom dataset designed to simulate real-world physical dynamics. Built a training and evaluation framework on both GPU and NPU using Megatron-LM, MindSpeed-LLM. Also took part in writing the paper.

Selftok: Discrete Visual Tokens of Autoregression Aug. 2024 - Present

- **Background**: Traditional Tokenizers often rely on spatial priors or continuous representations, which have limitations in effectively supporting reinforcement learning and integrating seamlessly with LLMs.
- **Method**: We propose Selftok. It addresses these challenges by leveraging the reverse diffusion process to create discrete AR tokens.

- **Results**: Our VLM based on Selftok tokens achieve both SOTA visual comprehension and generation performances.
- **My Involvement**: Conducted the development of an inference and evaluation pipeline for the tokenizer, enabling automatic evaluation and inference on both NPU and GPU. Participated in the design and implementation of part of ablation experiments. Explored and built an AR training framework for selftok tokens. Curated and constructed a large-scale high-resolution dataset for training.

National University of Singapore Apr. 2024 - Dec. 2024
Research Intern Singapore

A Benchmark of Multimodal Large Language Model

- **Background**: The evaluation method which assumes that better performance across various tasks implies stronger capabilities is not the best choice for evaluating MLLMs.
- **Method**: We propose a General-Level framework, to assess the performance and generality of MLLMs across five levels. Central to this framework is the concept of Synergy, which categorizes capabilities based on whether MLLMs preserve synergy across comprehension, generation, and multimodal interactions.
- **Results**: We present General-Bench, a massive-ever multimodal benchmark encompassing a broader spectrum of skills, modalities, formats, and capabilities, with over 700 tasks and 325,800 instances. The evaluation results, involving over 100 state-of-the-art MLLMs.
- **My Involvement**: Collected and curated 40 datasets spanning multiple areas of image comprehension, including image classification, image segmentation, and OCR-based Visual Question Answering (VQA), etc. Then conducted performance evaluations of these datasets on 10 state-of-the-art multimodal large language models, such as GPT-4o and Qwen2-VL-7B. Finally evaluated their performance according to the 5 levels we set, systematically categorizing and assigning each model to an appropriate performance level based on the evaluation results.

ACHIEVEMENTS

CUMCM, Provincial Second Prize *Summer 2022*

SKILLS

Language Mandarin(native), English(IELTS: 6.5, GRE: 330)
Programming Languages Python, MATLAB, R, C++