

Haoran (Jensen) Yuan

✉ haorany3425@gmail.com • 🌐 [haorany7](https://haorany7.github.io) • 🌐 Webpage: haorany7.github.io

RESEARCH INTERESTS

Efficient AI, with a focus on making **embodied AI generalizable, scalable, and efficient**: world-action models, representation learning, visuo-tactile learning, lossless speculative decoding, agentic robotics, reinforcement learning for world models, and recursive self-improvement (RSI).

EDUCATION

Doctor of Philosophy in Computer Engineering <i>University of Illinois Urbana-Champaign Advisor: Prof. Ismini Lourentzou</i>	Aug. 2026 – May 2028 (Expected) Champaign, IL
Master of Science in Computer Engineering <i>University of Illinois Urbana-Champaign</i>	Sep. 2024 – May 2026 Champaign, IL
Bachelor of Science in Electrical Engineering, Minor in Computer Science <i>University of Illinois Urbana-Champaign</i>	Sep. 2020 – Jun. 2024 Champaign, IL
Bachelor of Engineering in Computer Engineering <i>Zhejiang University</i>	Sep. 2020 – Jun. 2024 Hangzhou, China

PUBLICATIONS

* *equal contribution* † *project lead* ‡ *equal advising*

- [1] **DexTacWAM: A Visuo-Tactile World-Action Model for Dexterous Manipulation**
Haoran Yuan[†], Zekai Wang, Boning Shao, Haoran Lu, Trevor Darrell, Ismini Lourentzou[‡], Wei Zhan[‡]
arXiv, 2026. [[project](#)] [[arXiv:2609.24976](#)] [[code](#)]
- [2] **VTAM: Video-Tactile-Action Models for Complex Physical Interaction Beyond VLAs**
Haoran Yuan^{*†}, Weigang Yi^{*}, Zhenyu Zhang^{*}, Wendi Chen^{*}, Yuchen Mo, Jiashi Yin, Xinzhuo Li, Xiangyu Zeng, Chuan Wen, Cewu Lu, Katherine Driggs-Campbell, Ismini Lourentzou
arXiv, 2026. [[project](#)] [[arXiv:2603.23481](#)]
- [3] **JetSpec++: Foresight Parallel Drafting with Self Refinements**
Zhaoxiang Feng^{*}, Lanxiang Hu^{*}, **Haoran Yuan**, Yu-Yang Qian, Jihui Yang, Peng Zhao, Yibo Zhu, Tajana Rosing, Hao Zhang
arXiv, 2026. [[arXiv:2610.19792](#)] [[code](#)]
- [4] **JetSpec: Breaking the Scaling Ceiling of Speculative Decoding with Parallel Tree Drafting**
Lanxiang Hu, Zhaoxiang Feng, Yulun Wu, **Haoran Yuan**, Yujie Zhao, Yu-Yang Qian, Bojun Wang, Peng Zhao, Daxin Jiang, Yibo Zhu, Tajana Rosing, Hao Zhang
arXiv, 2026. [[project](#)] [[arXiv:2606.18394](#)] [[code](#)]

EXPERIENCE

PLAN Lab, University of Illinois Urbana-Champaign *Urbana, IL*
Graduate Researcher / Ph.D. Student – Advisor: Prof. Ismini Lourentzou Sep. 2025 – Present

VTAM: Video-Tactile-Action Model (*Project Lead*)

- Adapted a pretrained video transformer to touch **without tactile pretraining**, reaching **90%** success on real robots.
- Designed gated cross-attention that cut training loss spikes by **42%** and sped up convergence by **1.7×**.
- Added a force-aware diffusion regularizer to balance tactile-vision attention: **+80%** success over vision-only.

Berkeley AI Research (BAIR), UC Berkeley

Visiting Researcher – Advisors: *Prof. Wei Zhan, Prof. Ismini Lourentzou, Prof. Trevor Darrell*

Berkeley, CA

Apr. 2026 – Present

DexTacWAM: Dexterous Visuo-Tactile World-Action Model (*Project Lead*)

- Adapted a pretrained video DiT to 10 fingertip tactile sensors via **modality-split flow matching**.
- Designed a finger- and pose-aware tactile compressor: **89.4%** contact recall retained, **3.67×** fewer DiT tokens.
- Achieved **70.6%** avg. success on 6 real-robot tasks vs. **38.0%** for the best baseline; featured by BAIR.

Hao AI Lab, UC San Diego

Research Collaborator – Advisor: *Prof. Hao Zhang*

San Diego, CA

Feb. 2026 – Present

JetSpec++: Foresight Parallel Drafting with Self Refinements (*Core Contributor*)

- Developed a **causal correction for parallel drafting** that conditions each drafted token on its full branch history.
- Outperformed DeepSeek’s DSpark at matched size; up to **7.45×** lossless speedup over AR (**1,361 tok/s** on GB200).

JetSpec: Parallel Tree Drafting for Speculative Decoding (*Core Contributor*)

- Co-developed a **causal parallel draft head** that emits a complete speculative tree in one forward pass.
- Achieved up to **9.64×** lossless speedup over AR on Qwen3-8B, beating DDTree (8.78×) and EAGLE-3.

Snowflake

Software Engineering Intern

Menlo Park, CA

May 2025 – Aug. 2025

- Built a **predictive autoscaler** for GS clusters using Gradient Boosting, ARIMA and XGBoost models.
- Achieved **14.3% MAPE** and **92.5%** coverage (95% CI) for 24-hour CPU forecasts, in <5 min per deployment.
- Deployed to adjust instance provisioning, raising average **CPU utilization by >20%** in simulated datacenters.

TECHNICAL SKILLS

Languages: Python, C/C++, Rust, Java, SQL, SystemVerilog.

Frameworks: PyTorch, CUDA, Triton, SGLang, vLLM, Isaac Gym, Isaac Sim, MuJoCo, Flask, SQLAlchemy, OpenCV.

Databases & Tools: MySQL, PostgreSQL, Redis, Docker, AWS, GCP, Git, Linux, Shell.