Yuxuan Zhang

Coding Agents · Harness Engineering · Empirical Agent Evaluation

Vector Institute · PhD at UBC

+1-437-987-9608 · reacher.zh@gmail.com · **https://reacher-z.github.io** · Google Scholar · LinkedIn

Education

University of British Columbia — Jan. 2025 – Present
*PhD in Computer Science, advised by **Prof. Kelsey Allen** (Senior Research Scientist, **DeepMind**)* — *Vancouver, Canada*

University of Toronto — Sept. 2022 – June 2024
*MSc in Applied Computing, GPA: **4.0/4.0*** — *Toronto, Canada*

Peking University — Sept. 2019 – June 2022
BSc in Economics — *Beijing, China*

Publications

Underline = me; ⋆ = first author; ∗ = equal contribution.

2026 ⋆ Zhang, Y. et al. ***ClawBench: Can AI Agents Complete Everyday Online Tasks?*** *Under review.* [Project] [Paper]

2026 ⋆ Zhang, Y. et al. ***RewardHarness: Self-Evolving Agentic Post-Training.*** *Under review.* [Project]

2025 *ScholarCopilot: Training Large Language Models for Academic Writing with Accurate Citations. In Conference on Language Modeling (**COLM 2025**).* [Project] [Paper]

2025 *Retri3D: 3D Neural Graphics Representation Retrieval. In **ICLR 2025 (Spotlight, top 3%)**.* [Project] [Paper]

2025 *Dr. Bench: A Multidimensional Evaluation for Deep Research Agents, From Answers to Reports. Under review.* [Project] [Paper]

2025 *StructEval: Benchmarking LLMs' Capabilities to Generate Structural Outputs. In **TMLR 2025** (J2C Certification).* [Project] [Paper]

2025 ⋆ Zhang, Y. et al. *Watch Before You Answer: Learning from Visually Grounded Post-Training. In **CVPR 2026 Findings**.* [Project] [Paper]

2025 *PLAICraft: Large-Scale Time-Aligned Vision-Speech-Action Dataset for Embodied AI. In **CVPR 2026 Findings**.* [Project] [Paper]

2025 *VideoScore2: Think Before You Score in Generative Video Evaluation. Under review.* [Project] [Paper]

2025 *WikiGap: Promoting Epistemic Equity by Surfacing Knowledge Gaps Between English Wikipedia and Other Language Editions. Under review.* [Project] [Paper]

2024 Zhou, M.*, Zhang, Y.*, and Xu, X. *Edge-Enhanced Dilated Residual Attention Network for Multimodal Medical Image Fusion. In International Conference on Bioinformatics and Biomedicine (**BIBM 2024**).* [Project] [Paper]

2021 Chen, Y.*, Zhang, Y.*, Huang, Z., Luo, Z., and Chen, J. *CelebHair: A New Large-Scale Dataset for Hairstyle Recommendation Based on CelebA. In International Conference on Knowledge Science, Engineering and Management (**KSEM 2021**).* [Paper]

Research Experience

Research Assistant, ClawBench Jan. 2025 – Present

University of British Columbia, advised by Prof. Kelsey Allen, Vector Institute *Vancouver, Canada*

- Led **ClawBench**, an **open-source empirical agent evaluation and post-training infrastructure** for real-world browser agents, with **153 everyday online tasks** across **144 live platforms** and human reference traces for write-heavy workflows
- Built a safe live-web agent harness with Chrome/CDP instrumentation and final-request interception, preventing real-world side effects while converting production-site interactions into **five-layer traces** for evaluation and training
- Implemented an **Agentic Evaluator** that compares agent rollouts against human references, producing binary task-success verdicts with step-level justifications for debugging and future harness iteration
- Benchmarked **7 frontier models**, showing that the strongest model, Claude Sonnet 4.6, completed only **33.3%** of tasks while GPT-5.4 completed **6.5%**, exposing a deployment-reliability gap hidden by sandbox benchmarks

Research Assistant, VidGround Jan. 2025 – Present

University of British Columbia, advised by Prof. Kelsey Allen, Vector Institute *Vancouver, Canada*

- Exposed pervasive **linguistic shortcutting** in video understanding benchmarks and post-training datasets, showing that **40–60%** of questions in popular benchmarks (VideoMME, MMVU) could be answered correctly from text alone without any visual input
- Introduced **VidGround**, a data curation approach for post-training VLMs that excluded questions answerable from text alone and retained only visually grounded ones
- Demonstrated that training on only **69.1%** of the original post-training data with a simple RL-based algorithm improved video understanding performance by up to **6.2 points** on standard benchmarks
- Outperformed more complex RL post-training techniques including TW-GRPO, LongVILA-R1, and Video-RTS, identifying **data quality as the primary bottleneck** for VLM video understanding

Research Assistant, ScholarCopilot Oct. 2024 – Present

University of Waterloo, advised by Prof. Wenhu Chen *Waterloo, Canada*

- Framed citation hallucination as an AI reliability failure in scholarly writing, where LLMs confidently generate fabricated references indistinguishable from legitimate citations — undermining scientific integrity at scale
- Built **ScholarCopilot**, an agentic retrieval-augmented generation (RAG) framework that enables LLMs to write and cite simultaneously, addressing hallucination by retrieving real references *during* generation rather than post hoc
- Trained and evaluated the model on a 500k-paper arXiv corpus with aligned references, jointly optimizing next-token prediction loss and contrastive retrieval loss
- Achieved **40.1% top-1 citation recall**, outperforming **E5-Mistral-7B** (**15.0%**) and **BM25** (**9.8%**), while improving writing quality scores to **16.2/25**
- Validated in a user study where **all 10 participants preferred** ScholarCopilot's citations over ChatGPT, confirming reduced hallucination and stronger reference grounding

Research Assistant, PLAICraft Jan. 2025 – Dec. 2025

University of British Columbia, advised by Prof. Frank Wood *Vancouver, Canada*

- Co-developed **PLAICraft** (**CVPR 2026 Findings**), the first large-scale time-aligned vision-speech-action dataset for embodied AI, capturing **10,000+ hours** of socially interactive Minecraft gameplay from **10,000+ participants worldwide** — the largest open multimodal embodied dataset to date
- Built a rigorous embodied-agent evaluation harness covering object recognition, spatial reasoning, language grounding, and long-term memory, establishing the first standardized benchmark for vision-speech-action agents
- Studied model-based behavior learning from third-person observations, showing that an agent could predict another agent's future states and actions from limited observations and generalize across environments

Research Assistant, Retri3D Mar. 2024 – Dec. 2024

University of Toronto, advised by Prof. Nandita Vijaykumar *Toronto, Canada*

- Developed **Retri3D** (**ICLR 2025 Spotlight, top 3%**), the first retrieval framework for **3D Neural Graphics Representations** (NeRF, 3DGS) via open-vocabulary text queries, enabling scene-level semantic search without per-scene retraining
- Minimized storage and retrieval overhead: **20 MB** embeddings vs. 19 GB for LangSplat (∼**900**× **smaller**) and query-time retrieval of 5×10^{-5} s vs. 17 s for LERF (**340,000**× **faster**)
- Designed an artifact-aware rendering pipeline with an adaptive camera movement strategy, achieving **64.7%** retrieval accuracy on LERF with only **5 rendered images** — within 3.8% of the training-data upper bound
- Improved retrieval accuracy across multiple datasets (e.g., LERF and ScanNet++) by leveraging pre-trained Vision-Language Models (VLMs)
- Validated Retri3D across diverse NGR formats, achieving state-of-the-art retrieval accuracy and demonstrating generalization beyond the training distribution

Data Scientist & Research Intern May 2023 – Dec. 2024

SOTI Inc., Mitacs Accelerate Program, advised by Prof. Nandita Vijaykumar *Mississauga, Canada*

- Led the on-device perception workstream in a **team of 5** for SOTI's drone-based warehouse inventory product, shipping models that ran in production on Jetson edge hardware
- Designed and deployed a lightweight **language-embedded 3D representation** on Jetson, enabling real-time scene understanding and natural-language-guided object localization for drone-mounted cameras during autonomous stock scanning
- Optimized SegFormer for the same edge stack via inference-engine tuning and quantization, cutting inference latency by **66.56%**, reducing model size by **43.8%**, and improving **mIoU** by **16.14%** over the baseline
- Proposed a **feature-based knowledge distillation** method combining self-supervised learning with a reused teacher classifier, reaching **79.9%** on CIFAR-100 (ResNet-8x4) and surpassing prior SOTA (SimKD, DIST, DKD) by **1.83%**

Technical Skills

Agent Infrastructure	Browser automation, Chrome CDP, action tracing, HTTP tracing, agent scaffolds
Programming	Python, C/C++, JavaScript, Bash
Post-Training	PyTorch, Verl, TRL, LlamaFactory, GRPO/PPO, PEFT/LoRA, reward modeling
LLM Systems	Transformers, vLLM, DeepSpeed, Accelerate, CUDA, Slurm, Docker, AWS
Evaluation	Agent benchmarks, trajectory logging, HTTP trace analysis, multimodal evals

Service

Reviewer, **ICLRW 2026**, **COLM 2026**, **CVPR 2026**, **ECCV 2026**, **ACM MM 2025–2026**

Honors & Awards

Mitacs Accelerate Fellowship (**$30,000**, University of Toronto)	2023
Academic Excellence PhD Award (**$1,985**, UBC)	2025
First Prize in Mathematical Contest in Modeling (MCM)	2021
Outstanding Interviewer Award (**$1,000**, Peking University)	2020