Baihan Zhang

zhang.baihan02@gmail.com | linkedin.com/in/baihan-zhang-5b25032b2

Education

Stanford University Palo Alto, CA

B.A., M.S. in Computer Science

- **Activities:** Columnist at Stanford Blockchain Review, Member at Stanford Quantum and Stanford AI Alignment

Publications

Zhang, B., Li, P., & Boneh, D. (2026). "Indirect Prompt Injection Defenses and Datasets Suffer from Shortcuts." *Manuscript in preparation.*

Zhang, B., Aw, K. L., Liu, Y., Kotar, K., & Yamins, D. L. K. (2026). "Predict What to Predict Next: Recursive Self-Improvement Through Multimodal Streams of Thought." *Manuscript in preparation.*

Zhang, B., Liu, Y., Aw, K. L., & Yamins, D. L. K. (2026). "Counterfactual Flow Benchmark: Testing Physical Grounding in Video World Models." *Manuscript in preparation.*

Aw, K. L., Kotar, K., **Zhang, B.**, Liu, Y., Chen, L. N., Venkatesh, R. M., Kazemian, A., Lee, W., & Yamins, D. L. K. (2026). "VAST: Visual Abstractions as Streams of Thought." *Cognitive Computational Neuroscience (CCN) 2026 Extended Abstracts.*

Research

Snowflake Menlo Park, CA

Research Intern *June 2026 – Present*

- Led codescan harness optimization by designing an agentic "self-evolve" loop for a 40% drop in token spending after matching top performance on CyberGym at 90% recall. This project is on track to save Snowflake $19M of quarterly token cost.
- Deployed a cascading agentic architecture that matches frontier performance for internal agent trace security monitoring at 10% of token cost.

Stanford NeuroAI Lab, advised by Prof. Dan Yamins Stanford, CA

Research Assistant *Feb 2026 – Present*

- Led GRPO post-training that teaches a video-language model to reason with world-model tokens, yielding significant gains on out-of-distribution video-QA benchmarks.
- Scaled multimodal data generation to 2.7M+ video clips across a 10-node GPU cluster, feeding lab-wide training and evaluation.
- Owned evaluation across 10+ benchmarks and authored the team's first results journal.

Stanford Computer Security Group, advised by Prof. Dan Boneh Stanford, CA

Research Assistant (Building Defense Against Agentic Prompt Injections) *March – June 2026*

- First-authoring publication on a novel prompt injection defensive system trained simultaneously on agent internal activations through SAE and raw untrusted text that the agent interacts with, showing a 10x cost increase for adversarial training attacks.
- Built a robust benchmark of 16k examples that increased attack diversity by 800% compared to the best existing public benchmarks through a novel reinforcement learning red-teaming technique.

Projects

Website Optimizer for On-Chain Agents (YC Hackathon Top 10) | *React, AI Agents* Nov 2025

- Wrote a novel web format standard (commerce.txt) to structure e-commerce sites, optimized for LLM agent interaction and payment execution.
- Wrote general-purpose code to convert e-commerce sites into commerce.txt.
- Placed Top 3 out of 150 teams, selected to pitch directly to Y Combinator and Stripe partners.

Technical Skills

Languages: C++, Python (PyTorch), JavaScript, HTML/CSS, Solidity
Systems & Tools: Linux, Git, MongoDB, AWS, LaTeX, Figma