Xutao Mao

xutao.henry.mao@gmail.com · (+86) 173-1925-3326 · henrymao2004 · henrymao2004.github.io

Education

City University of Hong Kong, Hong Kong — Sep. 2026 – May 2030 (Expected)

Ph.D. student in Computer Science

Vanderbilt University, USA — Aug. 2022 – Apr. 2026

B.S. in Computer Science & Mathematics

Publications

Conference

- **Xutao Mao**, Xiang Zheng, Cong Wang, et al. STARE: Step-wise Temporal Alignment and Red-teaming Engine for Multi-modal Toxicity Attack. **ICML 2026** (Poster).
- **Xutao Mao**, et al. MindVote: When AI Meets the Wild West of Social Media Opinion. **AAAI 2026** (Oral).
- Tao Liu*, **Xutao Mao***, Hongying Zan, et al. LogicCat: Text-to-SQL Benchmark for Multi-Domain Reasoning Challenges. **AAAI 2026** (Poster). *(* = equal contribution, listed alphabetically)*

Selected Preprint & Under Review

- **Xutao Mao**, Cong Wang, et al. Agent Hacks Agent: Autoresearch for Production-Agent Red-Teaming.
- **Xutao Mao**, Cong Wang, Bo Han, et al. Taming CoT Obfuscation in VLMs: From Mechanistic Evidence to Activation-Level Enforcement.
- **Xutao Mao**, Bo Han, Cong Wang, et al. Agents Don't Just Agree, They Remember: Benchmarking Persistent Sycophancy in Stateful Personal Agents.

Research Experience

Agent Hacks Agent: Autoresearch for Production-Agent Red-Teaming [GitHub] Jun. 2026 – Present

- **Motivation:** Production tool-using agents change with every model and tool update, yet existing red-team tools optimize a judged attack-success score and keep surface artifacts (payloads, attack programs) that record *where* an attack landed but not *why* a trajectory turned unsafe — hard to audit, patch, or reuse once the setting changes.
- **Approach & Results:** Proposed **AHA**, an autoresearch red-team framework whose falsifiable discovery loop commits a vulnerability hypothesis and a falsifier before writing the attack, executes it in a sandbox, reflects on the trajectory, and promotes evidence-confirmed findings into a **Vulnerability Concept Graph (VCG)** — each concept an auditable unit (mechanism + enabling condition + falsifier + transfer prediction + evidence). Across Claude Code and Codex, three scenarios, and direct/indirect attacks, the frozen VCG deployed single-shot on held-out data beats the strongest frozen discovery baseline by **14.2 points**, and concepts transfer across victim models and scenarios.

PASB: A Persistent-Sycophancy Benchmark for Stateful Personal Agents [GitHub] May 2026 – Jul. 2026

- **Motivation:** Stateful personal agents keep long-term profiles, memories, and reusable skills, turning sycophancy from a response-level alignment problem into *write-time memory governance*: a user's biased claim can be committed to durable state and reused in later neutral tasks.
- **Approach & Results:** Proposed **PASB**, a 1,600-task benchmark that isolates a 5-turn persist stage from a 3-turn neutral-query stage so any contamination must pass through durable state. Crossing the *commit boundary* is the single largest downstream-failure jump in the benchmark (**+27 points**), driven by **attribution stripping** — "the user said X" loses its source, evidential status, and scope once written into state.

Taming CoT Obfuscation in VLMs: Mechanistic Analysis and Activation-Level Enforcement
Dec. 2025 – Mar. 2026

- **Motivation:** RL-trained VLMs exhibit CoT obfuscation — reward rises while reasoning monitorability drops — with little mechanistic explanation in prior work.
- **Approach & Results:** Proposed **TAME**, combining behavior-level feedback with SAE template-feature suppression under a dual training constraint, plus a dual-gate monitor and three-stage interpretability analysis (attention + gradient + SAE). Monitorability improves by about **60%** over GRPO and generalizes across model families; reveals a representational pathology in which RL drives template features into content representations, mid-upper-layer attention surges, and SAE precision halves.

STARE: A Red-teaming Engine for Multi-modal Toxicity Attacks Aug. 2025 – Nov. 2025

- **Motivation:** VLMs are vulnerable to toxic-continuation attacks, but existing red-teaming treats image generation as a black box and cannot localize when toxic semantics emerge during synthesis.
- **Approach & Results:** Proposed **STARE**, a hierarchical-RL red-team engine (high-level prompt editing + low-level T2I GRPO fine-tuning) with step-wise temporal attribution. Attack success rate exceeds baselines by about **68%** with strong transfer; reveals a temporal attack surface in which conceptual toxicity concentrates in early–mid semantic construction while detail toxicity appears in late refinement, informing stage-wise defense.

Skills

- Transformer architecture and mainstream LLM/VLM families (Qwen, DeepSeek) and core mechanisms (MoE, RoPE, FlashAttention).
- Parameter-efficient fine-tuning (LoRA family), SFT, and RL training (PPO, GRPO, DPO).
- Agent / tool-use paradigms and mechanistic-interpretability methods (SAE, circuit tracing).
- Distributed training & inference frameworks (DeepSpeed, FSDP, vLLM) and toolchains (verl); multi-node multi-GPU parallelism and standard evaluation protocols.