Zonghuan Xu

Fudan University, School of Mathematical Sciences | 2430XH10002@m.fudan.edu.cn | Homepage | GitHub | Google Scholar | arXiv

Education

Fudan University, Shanghai, China
B.S. in Mathematics and Applied Mathematics, Xianghui Plan, School of Mathematical Sciences, 2024-present
Expected graduation: 2028 | GPA: 3.85 / 4.00 | Credits completed: 87.5

Coursework, excluding political/ideological courses: Mathematics: Mathematical Analysis A I-II (H), Advanced Algebra I-II, Analytic Geometry, Classical Mathematical Thought I-II, Mathematical Modeling, Ordinary Differential Equations, Abstract Algebra, Probability (H), Mathematical Statistics, Selected Topics in Geometry and Topology, Real Analysis (H), Complex Analysis (H), Numerical Algebra and Optimization (H), Academic Frontiers Seminar I; **AI / CS:** C Programming, Artificial Intelligence Foundations, Mathematical Foundations of AI, Algorithms and Data Structures (H), Natural Language Processing; **Other:** Microeconomics, Practical Communicative English Speaking, Academic English for Science and Technology, Philosophy of Art and Aesthetic Issues, Mental Health and College Life, Military Skills, Health Fitness II, Badminton I-II.

Research Interests

I am broadly interested in building more capable and trustworthy AI systems by revisiting fundamental assumptions and developing new problem formulations.

Publications

Xu, Z., Li, J., Zhao, Y., Zheng, X., Ma, X., and Jiang, Y.-G. *DropVLA: An Action-Level Backdoor Attack on Vision-Language-Action Models.* arXiv:2510.10932v4, 2026. arXiv

Li, J., Zhao, Y., Zheng, X., **Xu, Z.**, Li, Y., Ma, X., and Jiang, Y.-G. *AttackVLA: Benchmarking Adversarial and Backdoor Attacks on Vision-Language-Action Models.* arXiv:2511.12149v1, 2025. arXiv

Xu, Z., Zheng, X., Wu, Y., and Ma, X. *Beyond Surface Judgments: Human-Grounded Risk Evaluation of LLM-Generated Disinformation.* arXiv:2604.06820v1, 2026. arXiv

Xu, Z. and Ma, X. *From Order to Distribution: A Spectral Characterization of Forgetting in Continual Learning.* arXiv:2604.13460v1, 2026. arXiv

Xu, Z., Ma, X., and Jiang, Y.-G. *Human Model: The Missing Piece Toward Trustworthy AGI.* Position paper, 2026. OpenReview

Research Experience

Action-Level Backdoors in Vision-Language-Action Models (code)

Fudan University, Institute of Trustworthy Embodied AI | Jul 2025-Mar 2026 | arXiv preprint; under review at IROS 2026

- **What I did:** Led the main technical pipeline, from action-level VLA-backdoor framing and threat-model design to poisoned-data construction, OpenVLA/OFT implementation, LIBERO simulation experiments, ablation/transfer analyses, and manuscript drafting.
- **Framing contribution:** Framed action-level VLA backdoors as reusable and composable malicious action primitives that expose fine-grained embodied safety risks.
- **Technical contribution:** Built poisoned-data construction, proposed action-window relabeling for chunked action sequences, OpenVLA OFT/LoRA fine-tuning with 4-bit double quantization, LIBERO evaluation, trigger-modality and appearance ablations, and cross-suite transfer analysis.

Human-Grounded Evaluation of LLM-Generated Disinformation

Fudan University, Institute of Trustworthy Embodied AI | Nov 2025-Apr 2026 | arXiv preprint; under review at EMNLP 2026

- **What I did:** Took primary responsibility for the project pipeline, reframing it from broad LLM fake-news generation to human-grounded proxy-validity measurement and executing construct/questionnaire design, prompt protocols, dataset construction, model calls, human-judge alignment, statistics, and primary drafting.
- **Framing contribution:** Framed the project as a measurement-validity problem: defined reader-facing disinformation risk as the central construct, separated it from factual correctness and generic harmfulness, and evaluated LLM judges as proxies against aligned human judgments.
- **Technical contribution:** Built an aligned benchmark of 290 deceptive articles, 2,043 paired human ratings from 392 participants, and eight frontier judges; audited judge-human alignment across overall scoring, item-level ordering, textual-signal dependence, and generation-side premise checks.

Spectral Theory of Forgetting in Continual Learning

Fudan University, Institute of Trustworthy Embodied AI | Mar 2026-Apr 2026 | arXiv preprint; under review at NeurIPS 2026 main conference

- **What I did:** Took primary responsibility for the theory project, formulating the order-to-distribution question and completing the mathematical development, theorem/proof work, asymptotic analysis, synthetic validation, and manuscript drafting.
- **Framing contribution:** Reframed forgetting from an order-dependent phenomenon over fixed task sequences to a distribution-governed process where tasks are sampled from an underlying task distribution.
- **Technical contribution:** Developed an exact loss-level spectral characterization in an overparameterized linear regime, deriving an operator identity, replacing coarse O(1/k)-style bounds with generic exponential rates that are sharp up to constants, and identifying the leading asymptotic term.

Human Models for Trustworthy AI

Fudan University, Institute of Trustworthy Embodied AI | Apr 2026-May 2026 | Position paper; under review at NeurIPS 2026 position paper track

- **What I did:** Led the position paper's framing and drafting, including conceptual synthesis, argument structure, the primary manuscript draft, and the data-infrastructure proposal around situated experience records.
- **Framing contribution:** Synthesized recurring human-related AI methods into a unified human-model framing and made the generalization of setting-specific human models the central research question.
- **Technical contribution:** Proposed situated experience records as data infrastructure for human-model training and inference, preserving task context, AI actions, human responses, explicit feedback, and downstream outcomes within the same activity and as persistent inference-time context.

Additional Projects

World-Model Inputs for Atari Policies (code)

- **Goal:** Tested on Atari whether a policy performs better when it receives extra predictions from a world model trained alongside the policy, in addition to raw game observations.
- **What I did:** Made the framework-level change needed to pass a detached world-model summary into the policy input, then tried simple comparison designs such as directly appending the summary or letting the policy learn a small gate over it. I also repaired training edge cases and ran the full 26-game x 2-seed Atari100K experiment.
- **Result:** Average performance was close to the baseline, with mean/median HNS at 1.875/0.959 vs. 1.818/0.773, but the effect was highly unstable: some games improved substantially over the baseline or reported SOTA reference, while others degraded substantially. Follow-up ablations and mechanism checks also varied sharply across configurations and seeds, so I did not claim a stable conclusion and temporarily set the project aside.

Fudan Sports Reservation Automation (code)

- **Goal:** Built a fully automated Fudan sports-venue reservation system: after the user sets the desired venue and time, it automatically books the slot, including recognizing the CAPTCHA and clicking the required characters.
- **What I did:** Crawled several thousand CAPTCHA images from the university site, labeled character examples, trained a lightweight single-character neural network, and combined those predictions with segmentation and ordering algorithms for idiom-style click CAPTCHAs.
- **Result:** The CPU-only pipeline finishes recognition and clicking in under 2 seconds and reports over 90% end-to-end recognition-click-submit success, with retry logs and screenshots for failed attempts.

Learned Iterative Refinement for Quadratic-Root Prediction (code)

- **Goal:** Explored whether neural networks can be framed as iterative numerical solvers by reusing a learned refinement block across multiple steps to predict the largest root of a quadratic equation.
- **What I did:** Built one-shot MLP/ResMLP baselines, iterative refinement models that feed the previous estimate back into later steps, and an RL variant that rewards error reduction.
- **Result:** One-shot models reached about 10^-2 to 10^-2.5 mean absolute error; iterative reuse did not reliably improve stability; PPO reached about 10^-4 for one-step prediction but degraded over longer refinement chains.

Technical Skills

Programming and research tooling: Python, PyTorch, NumPy, pandas, SciPy/statsmodels, Matplotlib/Seaborn, OpenCV, LaTeX, Git/GitHub, Linux/SSH, Bash/PowerShell, Conda/uv, Slurm job scripting, W&B.
ML / VLA / RL systems: Hugging Face Transformers/PEFT, OpenVLA/OFT/LoRA fine-tuning, bitsandbytes 4-bit loading, LIBERO evaluation, RLDS/TFDS dataset builders, trajectory conversion, action-window/chunk relabeling, PPO, Atari100K and world-model RL workflows.
LLM and data pipelines: OpenAI-compatible API pipelines for OpenAI/Claude/Gemini/Qwen-style models, prompt-template management, batch generation/judging, JSONL/CSV cleaning, human-survey design and alignment analysis, source-data/figure/table export.
Automation, analysis, and reproducibility: Playwright/Selenium browser automation, CAPTCHA segmentation/recognition, benchmark and poisoned-data construction, statistical checks, judge-human validity analysis, spectral/operator proof analysis, reproducibility packaging.

Research Funding

- Fudan Undergraduate Research funding / Junzheng Program project on backdoor attacks against VLA systems, 2025-2026.