Jingxuan Fan (She/Her/hers)

Boston, MA·jfan@g.harvard.edu·+1(909)344-4142
linkedin.com/in/jingxuan-fan/ · github.com/jingxuanf0214

EDUCATION

Harvard University **Cambridge, MA**
Ph.D. Candidate, Program in Neuroscience Expected Jan. 2026
Dissertation: Reinforcement Learning with Dense Intrinsic Rewards for Complex Environment Navigation
M.S. in Applied Math March 2024
Relevant Coursework: Reinforcement Learning, Neural Computation, Advanced Topics in Data Science, Physical Mathematics, TinyML and Efficient Deep Learning, Mathematical & Engineering Principles for Training Foundation Models

Massachusetts Institute of Technology **Cambridge, MA**
B.S. in Brain and Cognitive Science May 2020
Honors & Awards: Hans Lukas Teuber Award for Outstanding Academics, Walle J.H. Nauta Award for Outstanding Research

SKILL & INTERESTS

Programming skills: Python (PyTorch, Tensorflow, scikit-learn, Pandas, SciPy/NumPy), Matlab, SQL, Linux, Git
Modeling skills: LLM post-training, reinforcement learning, representation learning, multimodal foundation model

SELECTED RESEARCH EXPERIENCE

Harvard University ***Cambridge, MA***
PhD Researcher, Dept. of Neurobiology *Expected Jan. 2026*

- **Developed a visual-olfactory virtual reality environment** to study reward-guided spatial reasoning and cognition; **Designed and trained an RNN-PPO agent with internal curiosity module (RND)** to simulate behavioral policies for reward-guided navigation and foraging tasks; **Conducted state space analysis** to extract low-dimensional neural dynamics representations and investigated their role in composing observable parameters in behavioral policies (journal manuscript in prep)
- **Led efforts in** NBLAST-based neuron matching of *Drosophila* central brain across large scale connectome datasets; **Performed graph embedding and hierarchical clustering** to multi-level sensory-motor pathways
- **Developed a Mixture of Expert (MoE)** framework to combine diverse modalities and sources of genetic variants prediction evidence (functional evidence, protein language model prediction, population genetic evidence, clinical notes) to achieve SOTA performance on genetic variants cancer risk prediction

Master's Researcher, Dept. of Applied Mathematics *March 2024*

- **Trained PPO agents with state-dependent dense intrinsic rewards** for complex maze navigation tasks; characterized how dense rewards shape exploration behavioral policy and latent representations
- Developed an **entropy-penalized composition method** for multi-attribute reward models and demonstrated **improved results on reward model benchmarks** (submission to **COLM 2025**)
- **Developed** a framework to **generate large-scale synthetic rule pool** and **perform data-aware rule selection** for scoring preference data in the safety domain; Demonstrated **improved results on reward model benchmarks** using preference data scoring with the rule adaptor (**ICLR BiAlign 2025**)
- **Developed test prompts and trained SAE to dissect transformer circuits** for spatial relations generation in SOTA text-to-image diffusion models e.g. PixArt, SD3.5; applied feature steering to improve 2/3D spatial scores on T2ICompBench (submission to **CVPR Visual Concepts 2025**)
- **Developed an automated method** to generate a large-scale, domain-specific dataset of graduate-level applied mathematics problems; **Benchmarked** leading closed- and open-source LLMs on this dataset and **performed in-depth error analysis**; Developed a framework to improve this domain specific ability through **tool usage** and **finetuning** (**NeurIPS MATH-AI 2024, ICLR 2025)**

Massachusetts Institute of Technology ***Cambridge, MA***
Undergraduate Researcher, Picower Institute *Sept.2017-May 2020*

- Conducted smFISH, IHC, q-PCR and behavioral assays to study the neural circuit for danger signal detection and avoidance during social behaviors and co-authored a paper published in ***Nature***

Undergraduate Researcher, McGovern Institute *Sept.2018-May 2020*

- Designed single-nanometer iron oxide nanoparticles as dopamine-responsive MRI sensors, developed brain-wide delivery methods to assess its distribution and functionality; Co-authored two papers published in ***JACS*** and ***PNAS***

PROFESSIONAL EXPERIENCE

Harvard AI Safety Student Team, Technical Fellow *Feb. 2025-May 2025*
Meta, Research Intern *May 2024-Aug. 2024*

- Developed a novel image-based feature representation tailored to high density sEMG and used customized CV models for gesture decoding and input feature attributions
- Introduced manifold capacity as a theoretical metric for representation quality evaluation and multimodal SSL loss
- Implemented generative models to extract disentangled factors in sEMG for generalization and data augmentation

Axoft, Software Intern *Sept. 2023-Dec. 2023*

- Developed and maintained in-house software pipelines for fluorescence imaging processing and spike sorting
- Applied statistical and machine learning models for neural decoding from population spiking and LFP data

PUBLICATIONS

Li, X., Chen X., Fan, J, Gao, M., Jiang, H. (2025). Entropy-aware Attribute Composition of Multi-head Reward Models (https://arxiv.org/abs/2503.20995). Under review at **COLM 2025**

Li, X., Gao, M., Fan, J, Zhang, Z., Li, W. (2025). Data-adaptive Safety Rules for Training Reward Models (https://arxiv.org/pdf/2501.15453). **ICLR BiAlign 2025,** Under review at **ICML 2025**

Fan, J., Martinson, S., Wang, E.Y., Hausknecht, K. (2024). HARDMath: A Benchmark Dataset for Challenging Problems in Applied Mathematics (https://arxiv.org/pdf/2410.09988). **NeurIPS 2024 MATH-AI workshop, ICLR 2025**

Kwon, J.-T., Ryu, C., Lee, H., Sheffield, A., Fan, J., Cho, D. H., Bigler, S., Sullivan, H. A., Choe, H. K., Wickersham, I. R., Heiman, M., & Choi, G. B. (2021). An amygdala circuit that suppresses social engagement. **Nature**, 593(7857), 114–118.

Wei, H., Wiśniowska, A., Fan, J., Harvey, P., Li, Y., Wu, V., Hansen, E. C., Zhang, J., Kaul, M. G., Frey, A. M., Adam, G., Frenkel, A. I., Bawendi, M. G., & Jasanoff, A. (2021). Single-nanometer iron oxide nanoparticles as tissue-permeable MRI contrast agents. **Proceedings of the National Academy of Sciences**, 118(42).

Hsieh V., Okada S., Wei H., García-Álvarez I., Barandov A., Alvarado SR., Ohlendorf R., Fan J., Ortega A., Jasanoff A. (2019). Neurotransmitter-responsive nanosensors for T2-weighted magnetic resonance imaging. **Journal of the American Chemical Society**, 141 (40), 15751-15