Jiajun Fan

Portfolio: jiajunfan.com

Email: jiajunf3@illinois.edu
Mobile: 1-4479026670

Education

- **University of Illinois Urbana-Champaign** Urbana, IL, USA
  Ph.D. of Computer Science; GPA 4.0/4.0 *Aug. 2024 - May. 2029*
  - **Research Field**: Stable and Autonomous RL Post-Training for Large Generative Models (Flow/Diffusion, Multi-modal Reasoning LLMs).
  - **Service**: Reviewer at ICML 2025–2026, ICLR 2025–2026, NeurIPS 2024–2025, CVPR 2026, AAAI 2025, AISTATS 2025, KDD 2024.
  - **Relevant Course**: Machine Learning Algorithms for Large Language Models (LLMs).

- **Tsinghua University** Beijing, China
  M.Eng. of Computer Technology; GPA 3.97/4.0, Top 1.3% *Sept. 2021 - Jun. 2024*
  - **Service**: Reviewer at NeurIPS 2022–2023, ICML 2023–2024, ICLR 2024.
  - **Relevant Course**: Stochastic Processes (A). Big Data Systems (A^+). Digital Processing of Speech Signals (A). Data Visualization (A^+).

- **Nankai University** Tianjin, China
  B.Eng. of Intelligent Science and Technology; GPA 93.28/100 (3.9/4.0), 1/83 *Sept. 2017 - Jun. 2021*
  - **Honors**: Ranked 1st in major; National Scholarship twice (Top 1%).

Research Interest

- **Primary**: **Stable and autonomous RL** post-training for large generative models, spanning **flow/diffusion models** and **multi-modal reasoning LLMs**. Key focus on progressively reducing human intervention in the post-training pipeline—from online RL that eliminates human-collected data (ORW-CFM-W2), to adaptive divergence control for **collapse-free** training without manual KL tuning (ADRPO), to process rewards that remove reliance on human-annotated reasoning paths (CESAR), to fully autonomous **self-critique** where models become their own critics, eliminating hand-crafted rewards entirely.
- **Secondary**: **Scalable and stable** post-training methods that address critical challenges in RL-based fine-tuning: **collapse-free** training that preserves generation diversity throughout continuous optimization, adaptive exploration-exploitation trade-offs via self-regulated KL divergence, and effective **test-time scaling** that resolves inverse scaling in reasoning models—enabling superhuman-level audio reasoning capabilities.

Research Highlights

- **Self-Evolving RLHF for Flow Matching Generative Models**
  Sept. 2024 - Present
  - **Role**: **1)** Introduced a **self-evolving RLHF** framework (ORW-CFM-W2) that enables flow matching models to continuously optimize through online reward feedback **without relying on human-collected datasets** or likelihood calculations. **2)** Derived a tractable Wasserstein-2 distance bound for flow matching models, providing the **first theoretical guarantee** for collapse-free policy evolution. **3)** Established a unified perspective connecting flow matching fine-tuning with traditional KL-regularized RL, enabling controllable reward-diversity trade-offs.
  - **Achievements**: **1)** Achieved state-of-the-art alignment with orders of magnitude less data while maintaining generation diversity through theoretically-grounded regularization. **2)** Validated the framework's effectiveness by successfully fine-tuning large-scale models like Stable Diffusion 3 across challenging tasks including spatial understanding and compositional generation. 3) Published a paper called "Online Reward-Weighted Fine-Tuning of Flow Matching with Wasserstein Regularization" at **ICLR 2025** as first author.

- **Behavior Control in Reinforcement Learning**
  Sept. 2021 - Aug. 2024
  - **Role**: **1)** Introduced a unified framework called LBC to achieve behavior control in RL. **2)** Provided a unified perspective on diverse RL methods for behavior control and potential enhancements. **3)** Validated LBC's efficacy through rigorous theoretical support and extensive empirical experiments.
  - **Achievements**: **1)** Surpassed **24 human world records** and attained the pinnacle of performance among reinforcement learning algorithms across most tasks. **2)** Published a paper titled "Learnable Behavior Control: Eclipsing Human World Records in Atari Games through Sample-Efficient Behavior Selection" at **ICLR 2023** with **oral presentation** as first author.

- **Sample-Efficient Reinforcement Learning**
  Sept. 2020 - Aug. 2024
    - **Role**: **1)** Introduced a sample-efficient Reinforcement Learning (RL) framework known as GDI, which achieved human-level performance by optimizing the data distribution of RL agents. **2)** Supported the efficacy of GDI with a robust foundation, including both theoretical proofs and an extensive array of experiments conducted in Atari. **3)** Provided a unified perspective on various RL algorithms with GDI.
    - **Achievements**: **1)** Outperformed prior SOTA method Agent57 with **500x less data** and **twice** the average performance.**2)** Published a paper called "Generalized Data Distribution Iteration" at **ICML 2022** as first author.

Publications

1. **Fan, J.**, Ren, R., Li, J., Pandey, R., Shivakumar, P.G., Gu, Y., Gandhe, A., Liu, G., Bulyko, I. Incentivizing Consistent, Effective and Scalable Reasoning Capability in Audio LLMs via Reasoning Process Rewards. International Conference on Learning Representations 2026 (**ICLR 2026**).

2. Y Li, Y Meng, Z Sun, K Ji, C Tang, **J Fan**, et al. SP-VLA: A Joint Model Scheduling and Token Pruning Approach for VLA Model Acceleration. International Conference on Learning Representations 2026 (**ICLR 2026**).

3. Wang, Z., **Fan, J.**, et al. Variational Supervised Contrastive Learning. The Thirty-Ninth Annual Conference on Neural Information Processing Systems 2025 (**NeurIPS 2025**).

4. **Fan, J.**, et al. Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models. The Thirty-Ninth Annual Conference on Neural Information Processing Systems 2025 (**NeurIPS 2025**).

5. **Fan, J.**, et al. Online Reward-Weighted Fine-Tuning of Flow Matching with Wasserstein Regularization. International Conference on Learning Representations 2025 (**ICLR 2025**).

6. Ye Li, Chen Tang, Yuan Meng, **Jiajun Fan**, et al. PRANCE: Joint Token-Optimization and Structural Channel-Pruning for Adaptive ViT Inference. IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2025.

7. **Fan, J.**, et al. Efficient Design-and-Control Automation with Reinforcement Learning and Adaptive Exploration. **NeurIPS 2024** Workshop AI4Mat.

8. **Fan, J.**, et al. Learnable Behavior Control: Breaking Atari Human World Records via Sample-Efficient Behavior Selection. International Conference on Learning Representations 2023 (**ICLR 2023**), **oral presentation, ranked 5/4176**.

9. Hao Wang, Chen Zhichao, **Jiajun Fan**, et al. Optimal Transport for Treatment Effect Estimation. The Conference on Neural Information Processing Systems 2023 (**NeurIPS 2023**).

10. **Fan, J.**, Xiao, C. Generalized Data Distribution Iteration. International Conference on Machine Learning 2022 (**ICML 2022**).

11. Xiao, C., Shi, H., **Fan, J.**, & Deng, S. CASA: A Bridge Between Gradient of Policy Improvement and Policy Evaluation. In the proceedings of Deep Reinforcement Learning Workshop **NeurIPS 2022**, 2022.

12. **Fan, J**. A Review for Deep Reinforcement Learning in Atari: Benchmarks, Challenges, and Solutions. In the proceedings of **AAAI-22** Workshop on Reinforcement Learning in Games, 2021.

13. **Fan, J.**, Xiao, C., & Huang, Y. GDI: Rethinking What Makes Reinforcement Learning Different From Supervised Learning. In the proceedings of **AAAI-22** Workshop on Reinforcement Learning in Games, 2021.

14. Z Wang, **Jiajun Fan**, et al. ProteinZero: Self-Improving Protein Generation via Online Reinforcement Learning. Under Review, 2025.

15. Y Li, Y Meng, Z Sun, K Ji, C Tang, J Fan, et al. SP-VLA: A Joint Model Scheduling and Token Pruning Approach for VLA Model Acceleration. Under Review, 2025.

16. Wang E., Lian J., **Fan J.**, et al. Enhancing Sequential User Modeling with Large-kernel Convolution: A Lightweight Approach. The 30th SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, under review.

17. Wang E., Can Z., Yang Y., Pan L., **Fan J.**, et al. Unbiased Recommender Learning from Implicit Feedback: A Weak Supervision Perspective. The 30th SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, under review.

18. Wang E., **Fan J.**, et al. SPEC: Constructing Reliable Sequential User Model using Slide-window Spectrum. ACM TheWebConf 2024 Conference, under review.

19. E. Wang, H. Li, T. Liu, Y. Yang, **Fan J.**, X. Liu, and Z. Chen, “Unbiased recommender learning from implicit feedback: A progressive proximal transport approach,” in ACM TheWebConf 2024 Conference, under review.
20. Wang, H., Chen, Z., **Fan, J**., et al. Entire Space Counterfactual Learning: Tuning, Analytical Properties and Industrial Applications. The IEEE Transactions on Neural Networks and Learning Systems (**TNNLS**), under review.
21. Xiao, C., Shi, H., **Fan, J.**, & Deng, S. An Entropy Regularization Free Mechanism for Policy-based Reinforcement Learning. arXiv preprint arXiv:2106.00707.
22. **Fan, J.**, Ba, H., Guo, X., & Hao, J. Critic PI2: Master Continuous Planning via Policy Improvement with Path Integrals and Deep Actor-Critic Reinforcement Learning. arXiv preprint arXiv:2011.06752.

Patents

- Unified framework for model-free reinforcement learning algorithms Fan, J., Xiao, C. Unified framework for model-free reinforcement learning algorithms. CN112766497A[P].
- Hyperparameter tuning algorithm based on multi-arm gambling machine optimizer Fan, J. Hyperparameter tuning algorithm based on multi-arm gambling machine optimizer. CN112926629A[P].
- An unbiased estimation algorithm of behavior value function Fan, J., Xiao, C. An unbiased estimation algorithm of behavior value function. CN112926628A[P].
- Policy gradient algorithm based on double robust qualification trace Fan, J., Xiao, C. Policy gradient algorithm based on double robust qualification trace. CN112926735A[P].
- Asynchronous multi-arm gambling machine hyperparameter optimizer based on electoral college voting mechanism Fan, J. Asynchronous multi-arm gambling machine hyperparameter optimizer based on electoral college voting mechanism. CN112949850A[P].
- Real-time multi-hyperparameter controller Fan, J. Real-time multi-hyperparameter controller. CN113052252A[P].
- Hyperspace multi-coupling parameter optimizer based on multi-arm gambling machine combined with democratic voting Fan, J. Hyperspace multi-coupling parameter optimizer based on multi-arm gambling machine combined with democratic voting. CN113052253A[P].
- Fast and generalizable hyperspace coupling multi-parameter nonlinear optimizer Fan, J. Fast and generalizable hyperspace coupling multi-parameter nonlinear optimizer. CN113052248A[P].
- Reinforcement learning algorithm based on generalized combination strategy space Fan, J. Reinforcement learning algorithm based on generalized combination strategy space. CN113052312A[P].

Awards

• Outstanding Graduates (1%)	Tianjin, China, 2021
• Excellent Graduation Thesis of Nankai University	Tianjin, China, 2021
• Tang Lixin Scholarship (1%)	Tianjin, China, 2021
• **National Scholarship**, Nankai University (1/83)	Tianjin, China, 2020
• Nomination for Zhou Enlai Scholarship	Tianjin, China, 2020
• **National Scholarship**, Nankai University (1/83)	Tianjin, China, 2019
• 3rd Prize, Robocup@HOME Education World Final	Sydney, Australia, 2019
• Bronze Medal, **ACM / ICPC** Asia Regional Contest	Xuzhou, China, 2019
• National 2nd Prize, National College Students Mathematical Contest in Modeling (5%)	Tianjin, China, 2018
• The First Prize Scholarship, Nankai University (2/83)	Tianjin, China, 2018

Research Experience

- **RL Post-Training for Audio Large Language Models** Ph.D. Student
  UIUC, Urbana, IL, USA *May. 2025 - Present*
  - **Motivation**: Reasoning in Audio LLMs remains underexplored. Chain-of-thought reasoning often degrades inference performance—a phenomenon we term *test-time inverse scaling*—due to hallucinatory and inconsistent reasoning. Existing RL methods rely on outcome-only rewards or require laborious hand-crafted process rewards, failing to cultivate genuine reasoning capability.
  - **Results**: Proposed CESAR, an online RL framework using GRPO with multi-faceted reasoning process rewards that incentivize consistency, structured reasoning, and calibrated depth. Resolved test-time inverse scaling and achieved SOTA on MMAU Test-mini, outperforming Gemini 2.5 Pro and GPT-4o Audio.
  - **Publication**: "Incentivizing Consistent, Effective and Scalable Reasoning Capability in Audio LLMs via Reasoning Process Rewards" at **ICLR 2026** as first author.

- **RL Post-Training for Flow/Diffusion Generative Models** Ph.D. Student
  UIUC, Urbana, IL, USA *Aug. 2024 - May. 2025*
  - **Motivation**: Large generative models like diffusion and flow matching models lack the ability to continuously improve from reward feedback without catastrophic forgetting or mode collapse. Existing fine-tuning methods either require expensive human-collected datasets or fail to maintain generation diversity.
  - **Results**: Achieved state-of-the-art alignment performance with orders of magnitude less data. Successfully fine-tuned large-scale models including Stable Diffusion 3 on challenging tasks such as spatial understanding and compositional generation, while preserving generation diversity.
  - **Publication**: Published "Online Reward-Weighted Fine-Tuning of Flow Matching with Wasserstein Regularization" at **ICLR 2025** as first author. Published "Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models" at **NeurIPS 2025** as first author.

- **Reinforcement Learning for Robotics** Research Assistant
  Mila, Montreal, Canada *Oct. 2023 - May. 2024*
  - **Motivation**: Robot design and control are typically treated as separate problems, leading to suboptimal solutions. Jointly optimizing both in an end-to-end manner can significantly improve overall system performance.
  - **Results**: Combined the strengths of model-free RL and design-based RL, improving multiple metrics in MuJoCo control tasks. Introduced behavior control into design-based RL, improving sample efficiency.
  - **Publication**: Published at **NeurIPS 2024** Workshop AI4Mat.

- **Behavior Control in Reinforcement Learning** Research Assistant
  Shenzhen, China *Jun. 2021 - Oct. 2023*
  - **Motivation**: Existing RL algorithms struggle with complex tasks requiring diverse exploration strategies. A unified framework for controlling agent behavior through adaptive strategy selection was needed.
  - **Results**: Surpassed **24 human world records** in Atari games and achieved top performance among RL algorithms across most tasks, demonstrating exceptional sample efficiency.
  - **Publication**: Published "Learnable Behavior Control: Breaking Atari Human World Records via Sample-Efficient Behavior Selection" at **ICLR 2023** with **oral presentation** (ranked **5/4176**), as first author.

- **Sample-Efficient Reinforcement Learning (GDI)** Research Assistant
  Beijing, China *Sep. 2020 - Jun. 2021*
  - **Motivation**: State-of-the-art RL methods like Agent57 required enormous amounts of data to achieve human-level performance in Atari games. A more data-efficient paradigm was needed to make RL practical for real-world applications.
  - **Results**: Outperformed Agent57 with **500x less data** and **twice** the average performance on the Atari-57 benchmark, surpassing **22 human world records**.
  - **Publication**: Published "Generalized Data Distribution Iteration" at **ICML 2022** as first author.

- **Model-based Reinforcement Learning Algorithm** Research Assistant
  Tianjin, China *Apr. 2020 - Sep. 2020*
  - **Role**: Introduced a groundbreaking method for achieving unparalleled performance on the challenging MuJoCo environment, which represents a significant breakthrough in the field of reinforcement learning.
  - **Achievements**: Presented a cutting-edge paper titled "Critic PI2: Master Continuous Planning via Policy Improvement with Path Integrals and Deep Actor-Critic Reinforcement Learning."

- **Multifunctional Home Service Robot** Research Assistant
  Tianjin, China *Sep. 2018 - Jul. 2019*

  - **Role**: **1)** Devised a comprehensive solution for the entire project, along with a finite state machine diagram. **2)** Employed cutting-edge ROS-based algorithms for autonomous navigation and RRT path planning.
  - **Achievements**: Presented the project in the 2019 ROBOCUP Sydney World Finals and won **3rd Prize**.

Academic Service

• Reviewer of The 28th International Conference on Artificial Intelligence and Statistics	AISTATS 2025
• Reviewer of Forty-Second International Conference on Machine Learning	ICML 2025
• Reviewer of The Thirteenth International Conference on Learning Representations	ICLR 2025
• Reviewer of The 39th Annual AAAI Conference on Artificial Intelligence	AAAI 2025
• Reviewer of The Thirty-Eighth Annual Conference on Neural Information Processing Systems	NeurIPS 2024
• Reviewer of The 30th SIGKDD Conference on Knowledge Discovery and Data Mining	KDD 2024
• Reviewer of The Forty-first International Conference on Machine Learning	ICML 2024
• Reviewer of The Twelfth International Conference on Learning Representations	ICLR 2024
• Reviewer of The Thirty-seventh Annual Conference on Neural Information Processing Systems.	NeurIPS 2023
• Reviewer of The Fortieth International Conference on Machine Learning.	ICML 2023
• Reviewer of The Thirty-sixth Annual Conference on Neural Information Processing Systems.	NeurIPS 2022