Yifei Wang

32 Vassar St, Cambridge, MA 02139, USA
yifei_w@mit.edu • +1 (617) 352-0578 • https://yifeiwang77.com

WORKING EXPERIENCE

Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

- Postdoc, Computer Science and Artificial Intelligence Laboratory (CSAIL) — Dec 2023 – Present
  - Domain: Machine Learning, Self-supervised Learning, Foundation Models
  - Advisor: Prof. Stefanie Jegelka

EDUCATION

Peking University, Beijing, China

- Ph.D. in Applied Mathematics, School of Mathematical Sciences — Sep 2017 – Jul 2023
  - Thesis: Self-supervised Contrastive Learning: Theory and Method
  - Advisors: Prof. Yisen Wang, Prof. Zhouchen Lin, Prof. Jiansheng Yang

Peking University, Beijing, China

- B.S. in Data Science, School of Mathematical Sciences — Sep 2013 – Jul 2017
- B.A. in Philosophy (double degree), Department of Philosophy — Sep 2014 – Jul 2017

AWARDS & SCHOLARSHIPS

- Best Paper Award, ICML 2024 ICL Workshop — 2024
- Outstanding Ph.D. Dissertation Runner-Up Award, CAAI — 2024
  Awarded by Chinese Association for Artificial Intelligence (CAAI), the leading AI academic organization in China.
- Excellent Graduate of Beijing Municipality, Top 0.1%, Beijing — 2023
  Awarded for outstanding graduates among all Beijing universities.
- National Scholarship (twice), Top 0.1% nation-wide, China — 2021, 2022
- President Scholarship, Top 1% university-wide, Peking University — 2022
- Baidu Scholarship Nomination Award, Top 20 worldwide, Baidu Inc. — 2022
- Silver Best Paper Award, ICML 2021 AML Workshop — 2021
- Best Machine Learning Paper Award (1/685), ECML-PKDD — 2021

ROLES & RESPONSIBILITIES

- Area Chair, ICLR 2024, ICLR 2025 — 2024, 2025
- Organizer, NeurIPS 2024 Workshop on Red Teaming GenAI — 2024
- Organizer, MIT ML Tea Seminar — 2024
- Reviewer, NeurIPS, ICML, AISTATS, AAAI, LoG, ECML-PKDD, CVPR, ICCV, ACL — 2021 – 2024

RESEARCH INTERESTS

Machine Learning, Self-supervised Learning, Foundation Models, AI Safety, Interpretability

PUBLICATIONS

37 peer-reviewed publications. 25 as (co-)first author. * denotes shared first authorship.

REFEREED CONFERENCE AND JOURNAL PAPERS

[37] **Yifei Wang***, Yuyang Wu*, Zeming Wei, Stefanie Jegelka, Yisen Wang, A Theoretical Understanding of Self-Correction through In-context Alignment, in *Proceedings of the 38th Conference on Neural Information Processing Systems (**NeurIPS 2024**)*. **Best Paper Award at ICML 2024 ICL Workshop.**

[36] **Yifei Wang***, Kaiwen Hu*, Sharut Gupta, Ziyu Ye, Yisen Wang, Stefanie Jegelka, Understanding the Role of Equivariance in Self-supervised Learning, in *Proceedings of the 38th Conference on Neural Information Processing Systems (**NeurIPS 2024**)*.

[35] Sharut Gupta*, Chenyu Wang*, **Yifei Wang***, Tommi Jaakkola, Stefanie Jegelka, In-Context Symmetries: Self-Supervised Learning through Contextual World Models, in *Proceedings of the 38th Conference on Neural Information Processing Systems (**NeurIPS 2024**)*. **Oral Presentation (top 4) at NeurIPS 2024 SSL Workshop**.

[34] Xinyi Wu, Amir Ajorlou, **Yifei Wang**, Stefanie Jegelka, Ali Jadbabaie, On the Role of Attention Masks and LayerNorm in Transformers, in *Proceedings of the 38th Conference on Neural Information Processing Systems (**NeurIPS 2024**)*.

[33] George Ma*, **Yifei Wang***, Derek Lim, Stefanie Jegelka, Yisen Wang, A Canonization Perspective on Invariant and Equivariant Learning, in *Proceedings of the 38th Conference on Neural Information Processing Systems (**NeurIPS 2024**)*.

[32] Qixun Wang, **Yifei Wang**, Yisen Wang, Xianghua Ying, Dissecting the Failure of Invariant Learning on Graphs, in *Proceedings of the 38th Conference on Neural Information Processing Systems (**NeurIPS 2024**)*.

[31] Lin Li, **Yifei Wang**, Chawin Sitawarin, Michael W. Spratling, OODRobustBench: A Benchmark and Large-scale Analysis of Adversarial Robustness under Distribution Shift, in *Proceedings of the 41st International Conference on Machine Learning (**ICML 2024**)*, 2024.

[30] Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, **Yifei Wang**, Zeming Wei, On the Duality Between Sharpness-Aware Minimization and Adversarial Training, in *Proceedings of the 41st International Conference on Machine Learning (**ICML 2024**)*, 2024.

[29] Qi Zhang, Tianqi Du, Haotian Huang, **Yifei Wang**, Yisen Wang, Look Ahead or Look Around? A Theoretical Comparison Between Autoregressive and Masked Pretraining, in *Proceedings of the 41st International Conference on Machine Learning (**ICML 2024**)*, 2024.

[28] **Yifei Wang***, Qi Zhang*, Yaoyu Guo, Yisen Wang, Non-negative Contrastive Learning, in *Proceedings of the 12th International Conference on Learning Representations (**ICLR 2024**)*, 2024.

[27] **Yifei Wang***, Jizhe Zhang*, Yisen Wang, Do Generated Data Always Help Contrastive Learning?, in *Proceedings of the 12th International Conference on Learning Representations (**ICLR 2024**)*, 2024.

[26] Tianqi Du*, **Yifei Wang***, Yisen Wang, On the Role of Discrete Tokenization in Visual Representation Learning, in *Proceedings of the 12th International Conference on Learning Representations (**ICLR 2024**)*, 2024.

[25] Xiaojun Guo*, **Yifei Wang***, Zeming Wei, Yisen Wang, Architecture Matters: Uncovering Implicit Mechanisms in Graph Contrastive Learning, in *Proceedings of the 37th Conference on Neural Information Processing Systems (**NeurIPS 2023**)*, 2023.

[24] Qi Zhang*, **Yifei Wang***, Yisen Wang, Tri-contrastive Learning: Identifiable Representation Learning with Automatic Discovery of Feature Importance, in *Proceedings of the 37th Conference on Neural Information Processing Systems (**NeurIPS 2023**)*, 2023.

[23] **Yifei Wang***, Liangchen Li*, Yisen Wang, Balance, Imbalance, and Rebalance: Understanding Robust Overfitting from a Minimax Game Perspective, in *Proceedings of the 37th Conference on Neural Information Processing Systems (**NeurIPS 2023**)*, 2023.

[22] Ang Li*, **Yifei Wang***, Yisen Wang, Adversarial Examples Are Not Real Features, in *Proceedings of the 37th Conference on Neural Information Processing Systems (**NeurIPS 2023**)*, 2023.

[21] George Ma*, **Yifei Wang***, Yisen Wang, Laplacian Canonization: A Minimalist Approach to Sign and Basis Invariant Spectral Embedding, in *Proceedings of the 37th Conference on Neural Information Processing Systems (**NeurIPS 2023**)*, 2023.

[20] Qi Zhang*, **Yifei Wang***, Yisen Wang, On the Generalization of Multi-modal Contrastive Learning, in *Proceedings of the 40th International Conference on Machine Learning (**ICML 2023**)*, 2023.

[19] Jingyi Cui*, Weiran Huang*, **Yifei Wang***, Yisen Wang, Rethinking Weak Supervision in Helping Contrastive Representation Learning, in *Proceedings of the 40th International Conference on Machine Learning (**ICML 2023**)*, 2023.

[18] Zeming Wei, **Yifei Wang**, Yiwen Guo, Yisen Wang, CFA: Class-wise Calibrated Fair Adversarial Training, in *Proceedings of the IEEE / CVF Computer Vision and Pattern Recognition Conference (**CVPR 2023**)*, 2023.

[17] Qi Chen, **Yifei Wang**, Zhengyang Geng, Yisen Wang, Jiansheng Yang, Zhouchen Lin, Equilibrium Image Denoising with Implicit Differentiation, *IEEE Transactions on Image Processing (**TIP**)*, 32, 1868-1881, 2023.

[16] **Yifei Wang***, Qi Zhang*, Tianqi Du, Jiansheng Yang, Zhouchen Lin, Yisen Wang, A Message Passing Perspective on Learning Dynamics of Contrastive Learning, in *Proceedings of the 11th International Conference on Learning Representations (**ICLR 2023**)*, 2023.

[15] Zhijian Zhuo*, **Yifei Wang***, Yisen Wang, Towards a Unified Theoretical Understanding of Non-contrastive Learning via Rank Differential Mechanism, in *Proceedings of the 11th International Conference on Learning Representations (**ICLR 2023**)*, 2023.

[14] Rundong Luo*, **Yifei Wang***, Yisen Wang, Rethinking the Effect of Data Augmentation in Adversarial Contrastive Learning, in *Proceedings of the 11th International Conference on Learning Representations (**ICLR 2023**)*, 2023.

[13] Xiaojun Guo*, **Yifei Wang***, Tianqi Du*, Yisen Wang, ContraNorm: A Contrastive Learning Perspective on Oversmoothing and Beyond, in *Proceedings of the 11th International Conference on Learning Representations (**ICLR 2023**)*, 2023.

[12] Mingjie Li, **Yifei Wang**, Yisen Wang, Zhouchen Lin, Unbiased Stochastic Proximal Solver for Graph Neural Networks with Equilibrium States, in *Proceedings of the 11th International Conference on Learning Representations (**ICLR 2023**)*, 2023.

[11] Shiji Xin, **Yifei Wang**, Jingtong Su, Yisen Wang, On the Connection between Invariant Learning and Adversarial Training for OOD Generalization, in *Proceedings of the 37th AAAI Conference on Artificial Intelligence (**AAAI 2023**)*. **Oral Presentation**.

[10] Qi Zhang*, **Yifei Wang***, Yisen Wang, How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders, in *Proceedings of the 36th Conference on Neural Information Processing Systems (**NeurIPS 2022**)*. **Spotlight Presentation**.

[9] Qixun Wang*, **Yifei Wang***, Hong Zhu, Yisen Wang, Improving Out-of-distribution Robustness by Adversarial Training with Structured Priors, in *Proceedings of the 36th Conference on Neural Information Processing Systems (**NeurIPS 2022**)*. **Spotlight Presentation**.

[8] Yichuan Mo, Dongxian Wu, **Yifei Wang**, Yiwen Guo, Yisen Wang, When Adversarial Training Meets Vision Transformers: Recipes from Training to Architecture, in *Proceedings of the 36th Conference on Neural Information Processing Systems (**NeurIPS 2022**)*. **Spotlight Presentation**.

[7] Qi Chen, **Yifei Wang**, Yisen Wang, Zhouchen Lin, Optimization-induced Graph Implicit Nonlinear Diffusion, in *Proceedings of the 39th International Conference on Machine Learning (**ICML 2022**)*.

[6] Mingjie Li, Xiaojun Guo, **Yifei Wang**, Yisen Wang, Zhouchen Lin, G^2CN: Graph Gaussian Convolution Networks with Concentrated Graph Filters, in *Proceedings of the 39th International Conference on Machine Learning (**ICML 2022**)*.

[5] **Yifei Wang***, Qi Zhang*, Yisen Wang, Jiansheng Yang, Zhouchen Lin, Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap, in *Proceedings of the 10th International Conference on Learning Representations (**ICLR 2022**)*.

[4] **Yifei Wang**, Yisen Wang, Jiansheng Yang, Zhouchen Lin, A Unified Contrastive Energy-based Model for Understanding the Generative Ability of Adversarial Training, in *Proceedings of the 10th International Conference on Learning Representations (**ICLR 2022**)*. **Silver Best Paper at ICML 2021 AML Workshop**.

[3] **Yifei Wang**, Zhengyang Geng, Feng Jiang, Chuming Li, Yisen Wang, Jiansheng Yang, Zhouchen Lin, Residual Relaxation for Multi-view Representation Learning, in *Proceedings of the 35th Conference on Neural Information Processing Systems (**NeurIPS 2021**)*.

[2] **Yifei Wang**, Yisen Wang, Jiansheng Yang, Zhouchen Lin, Dissecting the Diffusion Process in Linear Graph Convolutional Networks, in *Proceedings of the 35th Conference on Neural Information Processing Systems (**NeurIPS 2021**)*.

[1] **Yifei Wang**, Yisen Wang, Jiansheng Yang, Zhouchen Lin, Reparameterized Sampling for Generative Adversarial Networks, in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (**ECML-PKDD 2021**)*. **Best Machine Learning Paper Award (1/685)**, invited to *Machine Learning*.

WORKSHOP PAPERS

[4] Ziyu Ye, Jiacheng Chen, Jonathan Light, **Yifei Wang**, Jiankai Sun, Mac Schwager, Philip Torr, Guohao Li, Yuxin Chen, Kaiyu Yang, Yisong Yue, Ziniu Hu. Reasoning in Reasoning: A Hierarchical Framework for Better and Faster Neural Theorem Proving. **NeurIPS 2024 Workshop** on Mathematical Reasoning and AI.

[3] Hanqi Yan, Yulan He, **Yifei Wang** (corresponding author). The Multi-faceted Monosemanticity in Multimodal Representations. **NeurIPS 2024 Workshop** on Responsibly Building the Next Generation of Multimodal Foundational Models.

[2] Lizhe Fang*, **Yifei Wang***, Khashayar Gatmiry, Lei Fang, Yisen Wang. Rethinking Invariance in In-context Learning. **ICML 2024 Workshop** on Theoretical Foundations of Foundation Models (TF2M).

[1] Jingyi Cui*, Weiran Huang*, **Yifei Wang**, Yisen Wang. AggNCE: Asymptotically Identifiable Contrastive Learning. **NeurIPS 2022 Workshop** on Self-supervised Learning. **Oral Presentation.**

PREPRINT

[4] Lizhe Fang*, **Yifei Wang***, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, Yisen Wang. What is Wrong with Perplexity for Long-context Language Modeling? arXiv:2410.23771 (2024).

[3] Qi Zhang*, **Yifei Wang***, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, Yisen Wang. Beyond Interpretability: The Gains of Feature Monosemanticity on Model Robustness. arXiv preprint arXiv:2410.21331 (2024).

[2] Qixun Wang, **Yifei Wang**, Yisen Wang, Xianghua Ying. Can In-context Learning Really Generalize to Out-of-distribution Tasks? arXiv preprint arXiv:2410.09695 (2024).

[1] Zeming Wei, **Yifei Wang**, Ang Li, Yichuan Mo, Yisen Wang . Jailbreak and guard aligned language models with only few in-context demonstrations. arXiv preprint arXiv:2310.06387 (2023). **Cited over 150 times and featured in Anthropic's research blog**.

INVITED TALKS

- A Principled Path to Safe Foundation Models, MIT ML Tea Seminar — Oct 2024
- Building Safe Foundation Models from Principled Understanding, New York University — Sep 2024
- Reimagining Self-supervised Learning with Context, Princeton University — Aug 2024
- Non-negative Contrastive Learning, Cohere AI — Jun 2024
- Self-supervised Learning of Identifiable Features, TU Munich — May 2024
- Non-negative Contrastive Learning, MIT LIDS Tea Seminar — Apr 2024
- Understanding and Applying Self-supervised Learning via Graph, Deep Potential — 2023
- Towards Theoretical Foundations of Self-Supervised Learning, KAIST — 2022
- Towards Truly Unlearnable Examples for Data Privacy, Chinese Academy of Science — 2022
- Reparameterized Sampling for GANs, Beijing Academy of Artificial Intelligence (BAAI) — 2021
- Reparameterized Sampling for GANs, Plenary Talk at ECML-PKDD 2021 — 2021

TEACHING EXPERIENCE

- Guest Lecturer, CSCI 3370: Deep Learning, Boston College — Fall 2024
  Instructor: Prof Yuan Yuan
- Teaching Assistant, Introduction to AI (Trustworthy ML Class) — Fall 2022
  Instructor: Prof Yisen Wang
- Teaching Assistant, Advanced Topics in Machine Learning — Fall 2022
  Instructor: Prof Yisen Wang
- Teaching Assistant, Advanced Mathematics — Spring 2021
  Instructor: Prof Chao Wang
- Teaching Assistant, Optimization Methods in Machine Learning — Fall 2019
  Instructor: Prof Zhouchen Lin
- Teaching Assistant and Co-instructor, Machine Learning — Fall 2017
  Instructor: Prof Tong Lin. I **instructed two-week classes** on Support Vector Machine.