

# Yifei Wang

32 Vassar St, Cambridge, MA 02139, USA  
yifei\_w@mit.edu • Google Scholar • <https://yifeiwang77.com>

## WORKING EXPERIENCE

**Massachusetts Institute of Technology (MIT)**, Cambridge, MA, USA  
▪ Postdoc, Computer Science and Artificial Intelligence Laboratory (CSAIL) Dec 2023 – Present  
• Advisor: Prof. Stefanie Jegelka

## EDUCATION

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

**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

## SELECTED HONORS & AWARDS

- Best Paper Award, ICML 2024 ICL Workshop 2024
- Silver Best Paper Award, ICML 2021 AML Workshop 2021
- Best Machine Learning Paper Award (sole, 1/685), ECML-PKDD 2021
- Outstanding Ph.D. Dissertation Runner-Up Award, CAAI 2024
- Excellent Graduate of Beijing Municipality 2023
- Excellent Graduate of Peking University 2023
- National Scholarship (twice) 2021, 2022
- President Scholarship at Peking University 2022

## RESEARCH INTERESTS

I am interested in developing efficient and robust algorithms for large-scale foundation models (generative models and discriminative models) and providing a deep understanding of the underlying principles.

## PUBLICATIONS

44 peer-reviewed publications (39 in NeurIPS, ICLR, and ICML); 28 as (co-)first author. >1k citations.  
\* denotes shared first authorship.

### Generative Models, Language Models, Self-correction, Reasoning

*I worked on understanding and improving the key functionalities and capabilities of generative models, including long-context understanding, self-correction, reasoning, and sampling.*

Lizhe Fang\*, Yifei Wang\* et al. **What is Wrong with Perplexity for Long-context Language Modeling?** in ICLR 2025.

- I proposed a long-context perplexity measure that emphasizes long-context relevant tokens at training and evaluation, improving the benchmark scores on LongBench, LongEval, and RULER by up to 22%.

Lizhe Fang\*, Yifei Wang\* et al. **Rethinking Invariance in In-context Learning** in ICLR 2025.

- We discovered an expressive invariant in-context learning scheme (InvICL) that achieves permutation invariance of in-context demonstrations while preserving autoregressive nature and full context awareness at the same time.

Chenheng Zhang, Mingqing Xiao, Yifei Wang et al. **Tool Decoding: A Plug-and-Play Approach to Enhancing Language Models for Tool Usage** in ICLR 2025.

- We proposed a simple training-free, plug-and-play constrained decoding scheme that significantly improves LLM performance at tool use (e.g. it rivals GPT-4o with a 7B model).

Yifei Wang, et al., **A Theoretical Understanding of Self-Correction through In-context Alignment**, in NeurIPS 2024. **Best Paper Award at ICML 2024 Workshop on In-context Learning.**

- I established the first rigorous understanding of LLMs' self-correction ability and develop a simple and efficient self-correction algorithm (CaC) that shows significant improvements across different tasks (e.g., BBQ, AdvBench).

Yifei Wang, et al., **A Unified Contrastive Energy-based Model for Understanding the Generative Ability of Adversarial Training**, in ICLR 2022. **Silver Best Paper at ICML 2021 AML Workshop.**

- I proposed to use adversarial learning as an alternative paradigm to maximum likelihood training of energy-based models (EBMs) and established its superior image generation quality on CIFAR-10.

Yifei Wang, et al., **Reparameterized Sampling for Generative Adversarial Networks**, in ECML-PKDD 2021. **Best Machine Learning Paper Award (1/685).**

- I developed a structure-aware MCMC sampling method for GANs that can leverage discriminators (akin to reward models) to guide and refine image generation at test time.

Xixun Wang, Yifei Wang, **Yisen Wang**, and Xianghua Ying. **Can In-context Learning Really Generalize to Out-of-distribution Tasks?** in **ICLR 2025**.

- With controlled experiments, we found that in-context learning still happens only with in-domain tasks and hardly generalizes to novel OOD tasks. In other words, LLMs' in-context abilities are learned essentially through training data with likewise tasks.

Xinyi Wu, Amir Ajorlou, **Yifei Wang**, Stefanie Jegelka, Ali Jadbabaie, **On the Role of Attention Masks and LayerNorm in Transformers**, in **NeurIPS 2024**.

Ziyu Ye, Jiacheng Chen, Jonathan Light, **Yifei Wang** et al. **Reasoning in Reasoning: A Hierarchical Framework for Better and Faster Neural Theorem Proving**. **NeurIPS 2024 Workshop on Mathematical Reasoning and AI**.

Qi Zhang, Tianqi Du, Haotian Huang, **Yifei Wang**, Yisen Wang, **Look Ahead or Look Around? A Theoretical Comparison Between Autoregressive and Masked Pretraining**, in **ICML 2024**.

## Self-supervised Learning, Unsupervised Representation Learning

*I led a coherent series of works for building principled understandings and algorithms for self-supervised representation learning (contrastive learning, masked autoencoders, multimodal learning, etc).*

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, **Yifei Wang**, and Yisen Wang. **Projection Head is Secretly an Information Bottleneck**. in **ICLR 2025**.

- We showed that projection heads serve as an information bottleneck that prevent features from collapsing toward the pretraining task (e.g. instance classification).

Sharut Gupta\*, Chenyu Wang\*, **Yifei Wang\***, Tommi Jaakkola, Stefanie Jegelka, **In-Context Symmetries: Self-Supervised Learning through Contextual World Models**, in **NeurIPS 2024**. **Oral Presentation (top 4) at NeurIPS 2024 SSL Workshop and featured by MIT**

**Yifei Wang\***, Kaiwen Hu\*, Sharut Gupta, Ziyu Ye, Yisen Wang, Stefanie Jegelka, **Understanding the Role of Equivariance in Self-supervised Learning**, in **NeurIPS 2024**.

**Yifei Wang\***, Jizhe Zhang\*, Yisen Wang, **Do Generated Data Always Help Contrastive Learning?**, in **ICLR 2024**.

Tianqi Du\*, **Yifei Wang\***, Yisen Wang, **On the Role of Discrete Tokenization in Visual Representation Learning**, in **ICLR 2024**. **Spotlight Representation**

Qi Zhang\*, **Yifei Wang\***, Yisen Wang, **On the Generalization of Multi-modal Contrastive Learning**, in **ICML 2023**.

Jingyi Cui\*, Weiran Huang\*, **Yifei Wang\***, Yisen Wang, **Rethinking Weak Supervision in Helping Contrastive Representation Learning**, in **ICML 2023**.

**Yifei Wang\***, Qi Zhang\*, Tianqi Du, Jiansheng Yang, Zhouchen Lin, Yisen Wang, **A Message Passing Perspective on Learning Dynamics of Contrastive Learning**, in **ICLR 2023**.

Zhijian Zhuo\*, **Yifei Wang\***, Yisen Wang, **Towards a Unified Theoretical Understanding of Non-contrastive Learning via Rank Differential Mechanism**, in **ICLR 2023**.

Qi Zhang\*, **Yifei Wang\***, Yisen Wang, **How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders**, in **NeurIPS 2022**. **Spotlight Presentation**.

**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 **ICLR 2021**.

**Yifei Wang**, Zhengyang Geng, Feng Jiang, Chuming Li, Yisen Wang, Jiansheng Yang, Zhouchen Lin, **Residual Relaxation for Multi-view Representation Learning**, in **NeurIPS 2021**.

## Algorithmic Robustness, Distribution Shifts, AI Safety

*I worked on principled algorithms to improve the robustness of foundation models against adversarial attacks and natural distribution shifts, during which I built SoTA robust SSL models (ICLR'23).*

Zeming Wei, **Yifei Wang** et al. **Jailbreak and guard aligned language models with only few in-context demonstrations**. arXiv:2310.06387. **Cited over 160 times. It was featured and scaled up in Anthropic's research blog, where it successfully jailbroke prominent LLMs including GPT and Claude**.

Xixun Wang, **Yifei Wang**, Yisen Wang, Xianghua Ying, **Dissecting the Failure of Invariant Learning on Graphs**, in **NeurIPS 2024**.

Lin Li, **Yifei Wang**, Chawin Sitawarin, Michael W. Spratling, **OODRobustBench: A Benchmark and Large-scale Analysis of Adversarial Robustness under Distribution Shift**, in **ICML 2024**.

Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, **Yifei Wang**, Zeming Wei, **On the Duality Between Sharpness-Aware Minimization and Adversarial Training**, in ICML 2024.

**Yifei Wang\***, Liangchen Li\*, Yisen Wang, **Balance, Imbalance, and Rebalance: Understanding Robust Overfitting from a Minimax Game Perspective**, in NeurIPS 2023.

Ang Li\*, **Yifei Wang\***, Yisen Wang, **Adversarial Examples Are Not Real Features**, in NeurIPS 2023.

Zeming Wei, **Yifei Wang**, Yiwen Guo, Yisen Wang, **CFA: Class-wise Calibrated Fair Adversarial Training**, in CVPR 2023.

Rundong Luo\*, **Yifei Wang\***, Yisen Wang, **Rethinking the Effect of Data Augmentation in Adversarial Contrastive Learning**, in ICLR 2023.

Shiji Xin, **Yifei Wang**, Jingtong Su, Yisen Wang, **On the Connection between Invariant Learning and Adversarial Training for OOD Generalization**, in AAAI 2023. *Oral Presentation.*

Qixun Wang\*, **Yifei Wang\***, Hong Zhu, Yisen Wang, **Improving Out-of-distribution Robustness by Adversarial Training with Structured Priors**, in NeurIPS 2022. *Spotlight Presentation.*

Yichuan Mo, Dongxian Wu, **Yifei Wang**, Yiwen Guo, Yisen Wang, **When Adversarial Training Meets Vision Transformers: Recipes from Training to Architecture**, in NeurIPS 2022. *Spotlight Presentation.*

## Interpretability

*I leveraged statistical perspectives (e.g., identifiability) to develop intrinsically interpretable foundation models (ICLR'24 and NeurIPS'23) and discover their practical benefits (ICLR'25, EMNLP'24, and NeurIPS-W'24).*

Qi Zhang\*, **Yifei Wang\***, et al. **Beyond Interpretability: The Gains of Feature Monosemanticity on Model Robustness**. in ICLR 2025.

- We show that feature monosemanticity brought by SAEs (extrinsic methods) and NCL (intrinsic methods) can significantly improve model robustness under multiple scenarios.

Hanqi Yan, Yanzheng Xiang, Guangyi Chen, **Yifei Wang**, Lin Gui, and Yulan He, **Encourage or Inhibit Monosemanticity? Revisit Monosemanticity from a Feature Decorrelation Perspective**, in EMNLP 2024.

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.

**Yifei Wang\***, Qi Zhang\*, Yaoyu Guo, Yisen Wang, **Non-negative Contrastive Learning**, in ICLR 2024.

Qi Zhang\*, **Yifei Wang\***, Yisen Wang, **Tri-contrastive Learning: Identifiable Representation Learning with Automatic Discovery of Feature Importance**, in NeurIPS 2023.

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

## Graph Representation Learning, Invariant and Equivariant Learning

*I contributed to several key methodologies in building more powerful, expressive, and robust graph representation learning algorithms, including diffusion process, spectral filtering, and canonicalization.*

George Ma\*, **Yifei Wang\***, Derek Lim, Stefanie Jegelka, Yisen Wang, **A Canonicalization Perspective on Invariant and Equivariant Learning**, in NeurIPS 2024.

George Ma\*, **Yifei Wang\***, Yisen Wang, **Laplacian Canonization: A Minimalist Approach to Sign and Basis Invariant Spectral Embedding**, in NeurIPS.

Xiaojun Guo\*, **Yifei Wang\***, Zeming Wei, Yisen Wang, **Architecture Matters: Uncovering Implicit Mechanisms in Graph Contrastive Learning**, in NeurIPS 2023.

Mingjie Li, **Yifei Wang**, Yisen Wang, Zhouchen Lin, **Unbiased Stochastic Proximal Solver for Graph Neural Networks with Equilibrium States**, in ICLR 2023.

Qi Chen, **Yifei Wang**, Yisen Wang, Zhouchen Lin, **Optimization-induced Graph Implicit Nonlinear Diffusion**, in ICML 2022.

Mingjie Li, Xiaojun Guo, **Yifei Wang**, Yisen Wang, Zhouchen Lin, **G<sup>2</sup>CN: Graph Gaussian Convolution Networks with Concentrated Graph Filters**, in ICML 2022.

**Yifei Wang**, Yisen Wang, Jiansheng Yang, Zhouchen Lin, **Dissecting the Diffusion Process in Linear Graph Convolutional Networks**, in NeurIPS 2021.

<b>SKILLS</b>	<ul style="list-style-type: none"> <li>▪ Programming Languages: Python (proficient), C, MATLAB, R</li> <li>▪ Deep Learning Frameworks: PyTorch (proficient), TensorFlow, JAX, Keras</li> <li>▪ Distributed Training: Extensive experience with multi-node, multi-GPU training using up to 64 A100 GPUs, the PyTorch Distributed Data Parallel (DDP) framework, and the Slurm platform.</li> </ul>	
<b>INVITED TALKS</b>	<ul style="list-style-type: none"> <li>▪ Principles of Foundation Models, CDS at New York University</li> <li>▪ Principles of Foundation Models, Boston University</li> </ul>	<p>Feb 2025 Jan 2025</p>
	Principles of Foundation Models, John Hopkins University	Dec 2024
	<ul style="list-style-type: none"> <li>▪ Towards Test-time Self-supervised Learning, Guest Lecture at Boston College</li> <li>▪ A Principled Path to Safe Foundation Models, MIT ML Tea Seminar</li> <li>▪ Building Safe Foundation Models from Principled Understanding, New York University</li> <li>▪ Reimagining Self-supervised Learning with Context, Princeton University</li> <li>▪ Non-negative Contrastive Learning, Cohere AI</li> <li>▪ Self-supervised Learning of Identifiable Features, TU Munich</li> <li>▪ Non-negative Contrastive Learning, MIT LIDS Tea Seminar</li> <li>▪ Understanding and Applying Self-supervised Learning via Graph, Deep Potential</li> <li>▪ Towards Theoretical Foundations of Self-Supervised Learning, KAIST</li> <li>▪ Towards Truly Unlearnable Examples for Data Privacy, Chinese Academy of Science</li> <li>▪ Reparameterized Sampling for GANs, Beijing Academy of Artificial Intelligence (BAAI)</li> <li>▪ Reparameterized Sampling for GANs, Plenary Talk at ECML-PKDD 2021</li> </ul>	<p>Nov 2024 Oct 2024 Sep 2024 Aug 2024 Jun 2024 May 2024 Apr 2024 2023 2022 2022 2021 2021</p>
<b>PROFESSIONAL SERVICE</b>	<ul style="list-style-type: none"> <li>▪ Area Chair, ICLR 2024, ICLR 2025</li> <li>▪ Organizer, NeurIPS 2024 Workshop on Red Teaming GenAI</li> <li>▪ Organizer, MIT ML Tea Seminar</li> <li>▪ Reviewer, NeurIPS, ICML, AISTATS, AAAI, LoG, ECML-PKDD, CVPR, ICCV, ACL</li> </ul>	<p>2024, 2025 2024 2024 2021 – 2024</p>