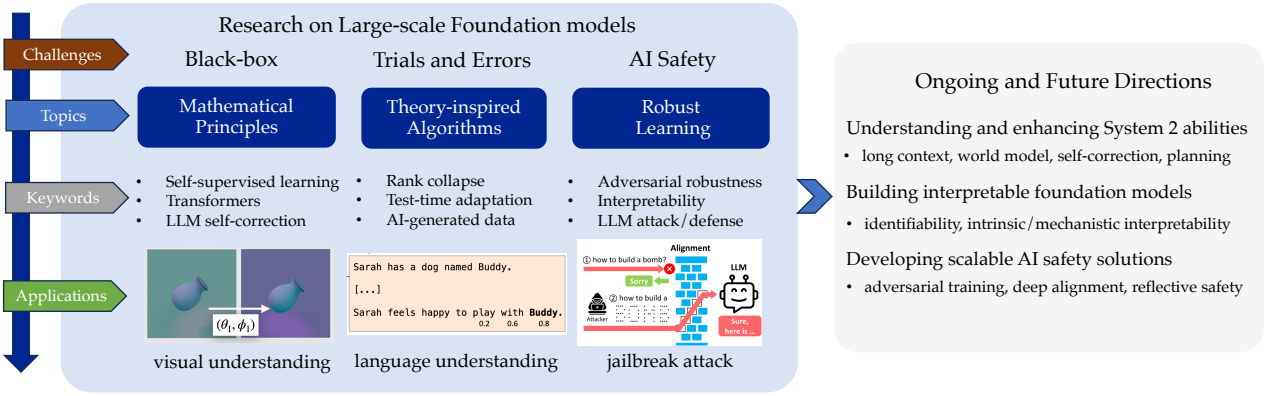
Principles of Large-scale Foundation Models

Yifei Wang, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology



Large-scale foundation models brought transformative paradigm shift in machine learning. However, their development remains predominantly empirical, leading to limited interpretability, costly trial and error, and unreliable behaviors. Addressing these challenges, **my research has contributed to uncovering the principles of foundation models and designing theory-inspired algorithms to enhance model capabilities and safety**.

1. **Mathematical Principles of Foundation Models**: I developed a coherent body of work that established theoretical foundations—covering generalization, training dynamics, and identifiability analyses—for a range of **self-supervised learning (SSL)** paradigms of foundation models. These include **autoregressive** [30], **reconstructive** [12, 22], **contrastive** [4, 6], **non-contrastive** [14], **predictive** [11] approaches, where I unified and characterized them within a graph-theoretic framework. For backbone networks like **Transformers**, I proposed dynamic analyses on its feature propagation [2, 19, 16, 29]. From an in-context learning perspective, I pioneered the first theoretical explanation of **LLMs' self-correction** ability (critical for **test-time reasoning** as in OpenAI o1 ↗), winning **Best Paper Award** at ICML'24 ICL workshop.

2. **Principled Algorithms for Enhancing Model Capabilities**: Building on these theoretical insights, I addressed practical challenges in foundation models, such as the **rank collapse** issue in **Transformers** [14, 19, 16, 12]. I developed novel training paradigms that rigorously enable **monosemantic** representations [7, 20] and **unsupervised test-time adaptation** [25] in foundation models (to be featured in **MIT CSAIL News**↗). To further scale foundation models with **AI-generated data**, I investigated their impact on generalization bias and proposed adaptive training strategies to mitigate this bias, achieving significant improvements on benchmarks [8]. Additionally, I developed efficient sampling algorithms that provably enhance the quality of generated data, earning the sole **Best Paper Award** at ECML'21 [1].

3. **Principled Robust Learning for Trustworthy Foundation Models**: The theoretical insights further inspired principled ways to build trustworthy foundation models in terms of adversarial robustness [9, 15, 21], interpretability [7, 20], and domain generalization [13, 27, 26]. For LLMs, I led a series of principled approaches that **use their own emergent abilities to address LLM safety**. Notably, our in-context jailbreak (cited over 150 times a year) [28] was featured and scaled up by **Anthropic**—the leading effort in AI safety today—showing in their blog ↗ that it can jailbreak many powerful LLMs (including GPT and Claude). Prior to **OpenAI o1** ↗, we were also the first to use *LLMs' self-correction ability* for jailbreak defense and showed superior performance to human-designed methods [10].

Impact: My research emphasizes a dynamic loop between theory and practice in large-scale foundation models, and it found successful applications in vision [4, 12], language [30, 10], graph [16, 19], and multi-modal [17] domains. My first-author papers have won **3 best paper awards** at ECML'21 [1] and two ICML workshops [5, 10], and 3 **spotlight presentations** at ICLR and NeurIPS. My thesis won **CAAI Outstanding Ph.D. Dissertation Runner-Up Award**. I was invited to share our findings with over ten institutes worldwide in industry and academia, including Princeton, NYU, TU Munich, and Cohere AI. My research has significantly influenced the understanding and design of many SSL methods and are frequently cited by top researchers.

Onging and Future Work: My long-term research goal is to build intelligent and reliable machines capable of solving complex tasks that are challenging even for humans. To achieve this, I aim to understand **the power and the limit of foundation models for achieving higher-level capabilities**, especially, planning and reflective decision-making [10], long-context understanding [35], and cross-modal reasoning [31]. My experience in SSL also supports me to explore **new self-supervised learning paradigms** beyond autoregressive models—as a preliminary step, I built joint embedding world models [25]. For reliable deployment, I have led the development of **intrinsically interpretable** ("white-box") models [20, 7] and **robust algorithms** that generalize to out-of-distribution domains [36, 37]. Additionally, I am interested in interdisciplinary studies that apply foundation model principles to accelerate **scientific discovery**. Currently, I am leading a collaboration at MIT that utilizes self-supervised learning for discovering influential factors from unlabeled **ocean dynamics** data.

Mathematical Principles of Foundation Models

The development of foundation models is driven by empirical research without solid understanding, while existing learning theories focused on classic problem setups often provide limited practical insights. Developing rigorous theories for these new foundation models will force us to delve into the essence of empirical designs, understanding their mechanisms and limitations. This will reduce trial and error, identify potential failures, and foster new algorithms. I developed new mathematical perspectives for understanding foundation models across different pretraining paradigms, backbone networks, and test-time reasoning abilities.

1.1 Understanding and unifying different SSL paradigms for pretraining: Foundation models rely on pretraining from massive unlabeled data, where two major SSL paradigms are **generative models** (predicting the input itself, e.g., next word prediction in GPT) and **joint embedding models** (matching samples in the embedding space, e.g., DINO and CLIP). By introducing a new connection between graph theory and SSL, I proposed a *graph-theoretic* formulation that unifies different types of SSL paradigms [4, 12, 17, 22, 30]. Specifically, I prove that every self-supervision (next word, masking, augmentation) *implicitly* induces a corresponding *similarity measure* $s(\cdot, \cdot)$ in the input space, which collectively constructs a *similarity graph* among all input samples. Based on this insight, I establish *generalization guarantees* based on spectral properties of the similarity graph, e.g., its algebraic connectivity. Along this line, we theoretically explained the design of masked autoencoders [12] and visual tokenizer [22], recognized as a NeurIP'22 **Spotlight paper** and an ICLR'24 **Spotlight paper**, respectively.

Apart from graph theory perspective, I also leverage diverse mathematical tools, e.g. *spectral analysis* [14], *information theory* [3], *probabilistic modeling* [5], *causal inference* [11], to analyze different SSL paradigms, offering fine-grained characterization tailed down to their uniqueness.

1.2 Understanding Transformers and LLM test-time reasoning: I develop dynamical system perspectives to understand the feature propagation of backbone networks in foundation models, including Transformers [16, 29], graph neural networks [2, 19], and asymmetric SSL designs [14]. In [19], I revealed the implicit bias of backbone networks during pretraining and showed how this synergy effect explains and predicts some unexpected model behaviors. In [29], we analyzed the implicit roles of attention mask and LayerNorm at preventing rank collapse in Transformers. Recently, I theoretically analyzed the mechanism of LLM self-correction and proved

that Transformer-based LLMs can refine responses through reflection on prior outputs. This work formalized self-correction as an *in-context alignment* process, proving how Transformers can refine predictions through reflective reasoning. This work pioneered the first rigorous theory for understanding self-correction in LLMs.

Principled Algorithms for Improving Model Capabilities

Inspired by the mathematical principles above, I develop practical algorithms that address model failures (such as rank collapse) and extend model capabilities (test-time adaptation). From a data-centric perspective, I also pioneered the study of using AI-generated data for further scaling foundation models.

2.1 Addressing key limitations limitations in foundation models and learning paradigms. I proposed a series of theory-driven solutions to the rank collapse of features that widely appear in many foundation models [14, 19, 16, 12]. From a contrastive learning perspective, I derived a normalization layer *ContraNorm* that improves Vision Transformers' accuracy by 5% [16]. For MAE, I addressed its rank collapse issue with a training regularization that improves MAE from 62.2% to 65.8% accuracy on ImageNet [12]. To avoid over-reliance on fixed data augmentations in previous SSL, I designed ContextSSL [25] that firstly enabled *test-time adaptation* of visual representations with a Transformer-based world model (**Oral** at NeurIPS'24 SSL workshop).

2.2 Scaling model abilities with AI-generated data: AI-generated data are being recognized as a valuable source for model training beyond human-generated data. Based on the theoretical principles above, I revealed the fundamental bias of synthetic data to SSL generalization and proposed a principled adaptive training strategy that yields substantial benefits [8]. I also explored using reward models to bootstrap the quality of synthetic data through tailored MCMC sampling, which not only theoretically guarantees convergence to the real data distribution but also improves sample efficiency a lot in practice (**ECML'21 Best Paper Award**) [1].

Principled Robust Learning for Trustworthy Foundation Models

Despite powerful, foundation models often fail catastrophically under unseen domains and adversarial inputs; the latter becomes a major threat to AI safety. My past research proposed principled robust learning algorithms against 1) real-world distribution shifts, 2) adversarial attacks, and 3) newly arising AI safety issues in LLMs.

3.1 New methodologies for OOD robustness. I pioneered two new methodologies in improving out-of-distribution (OOD) robustness. The first is the approach of **Canonicalization** [23, 24] for invariant and equivariant learning on structural data (e.g., graphs), which is model-agnostic and requires only preprocessing inputs, saving 41% training time compared to invariant networks. The second is to use **structural adversarial training** to address OOD generalization in [13, 27]. Additionally, I built a large-scale benchmark OODRobustBench [26] (with joint effort from UCB and KCL), which offers a comprehensive evaluation of OOD generalization of adversarial robustness with 706 robust models and 29 types of distribution shifts. It fits the *scaling law of OOD adversarial robustness* and alerts that existing methods are unlikely scaled to high OOD robustness.

3.2 Improved understandings and strategies for adversarial training: Adversarial Training (AT) is a fundamental solution to improve robustness to adversarial attacks. My research tackled several key problems in AT: 1) what are adversarial examples, 2) how to perform AT without labels, and 3) why AT suffers from robust overfitting. First, I challenged the classic understanding of "adversarial examples are features" by revealing that adversarial features cannot transfer between different learning paradigms so they are not "real features" [21]. Second, I identified the critical dilemma in self-supervised AT through a data augmentation perspective, resolving which improves robust accuracy by **10%** under AutoAttack [15]—a significant leap in a field where 1% gains are notable. Lastly, I developed a holistic explanation and solution to AT's long-standing robust overfitting problem through a minimax game perspective [9].

3.3 Scalable measures for improving AI safety in LLMs: The jailbreak attacks of LLMs raise most concerns where LLMs are manipulated to respond directly to harmful queries such as *"how to make a bomb"*. Different from existing work using costly gradient-based attack, I pioneer principled approaches to develop **scalable measures** for LLM safety based on *LLMs' own emergent abilities*. I first discovered that in-context learning (ICL) — LLMs' core emergent ability — can be used for crafting jailbreak attacks with few-shot (e.g. 5) harmful question-answer demonstrations. Anthropic features our work in their blog and scales it to jailbreak prominent LLMs (including GPT and Claude) with up to 256 shots [28]. We also show that in-context learning with a few safe demonstrations can also, in turn, improve LLM safety. Going further, we first propose to use *LLM's own self-reflection* – a core reasoning ability in empowering GPT-o1 – as a strong strategy for defending LLM jailbreaks and outperform many human-designed safety measures in practice [10].

Ongoing and Future Work

My ongoing and future work aims to develop **strong, interpretable, and robust** AI models that could learn efficiently from unlabeled and few-shot data and generalize to novel tasks. My previous research laid a solid foundation to seek possible breakthroughs through the following directions.

Understanding and enhancing *System 2* abilities: Addressing complex problems often requires long-range, reflective, and logical reasoning, commonly referred to as *System 2 thinking*, a capability increasingly evident in models like OpenAI's o1. I am interested in developing systematic understandings on the strengths and limitations of foundation models in *System 2* abilities and developing principled algorithms to enhance these capabilities. To improve models' understanding and reasoning abilities, I introduced novel perplexity metrics for long-context training [35] and developed contextual world models that adapt visual representations to new tasks at test time [25]. Additionally, I theoretically explained the self-correction mechanisms of LLMs [10] and devised hierarchical planning strategies [32], achieving significant gains in theorem proving. These advancements pave the way for models with more robust and reliable reasoning skills.

Building *interpretable* foundation models: I aim to develop foundation models that not only deliver accurate predictions but also ensure transparency, controllability, and trustworthiness. My work has focused on creating *intrinsically interpretable* models grounded in statistical principles [7, 20]. For example, I developed NCL [7], which guarantees theoretical feature identifiability and produces highly semantically coherent features, improving interpretability scores by 10× on ImageNet-100 without compromising performance. Furthermore, I have employed *mechanistic interpretability* tools to explore the vision-language modality gap, providing valuable *cognitive* insights aligned with human understanding [33]. In addition, I am leading a collaboration with MIT colleagues in *ocean science*, where we apply the NCL algorithm to uncover novel factors influencing ocean dynamics. This interdisciplinary effort highlights the broad applicability and impact of interpretable models in advancing scientific discovery and I look forward to more broad collaboration in this direction.

Developing *scalable* AI safety solutions: Current safe alignment techniques are often superficial and are easily circumvented. I advocate for the development of principled and scalable safety measures that enable AI capability and safety to *co-evolve*. My extensive experience in adversarial training [9, 5, 21, 26] equips me to design principled and efficient adversarial training methods tailored to foundation models. Similarly, my work in robust representation learning [13, 15, 27] provides a foundation for advancing from shallow alignment to *deep alignment*, shifting the focus from the output space to the *representation space*. Furthermore, I believe that systematically incorporating reflective thinking can pave a scalable path toward enhancing model safety (*reflective safety*). My research on in-context attack and defense [28] and self-correction-based jailbreak defenses [10] demonstrates significant potential in realizing this vision.

Reference (* denotes shared first authorship)

[1] **Yifei Wang**, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Reparameterized Sampling for Generative Adversarial Networks. In *ECML-PKDD*. 2021. ***Best ML Paper Award (1/685)***.

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

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

[4] **Yifei Wang***, Qi Zhang*, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap. In *ICLR*. 2022.

[5] **Yifei Wang**, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. A Unified Contrastive Energy-based Model for Understanding the Generative Ability of Adversarial Training. In *ICLR*. 2022. ***Silver Best Paper Award @ ICML 2021 AdvML Workshop***.

[6] **Yifei Wang***, Qi Zhang*, Tianqi Du, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. A Message Passing Perspective on Learning Dynamics of Contrastive Learning. In *ICLR*. 2023.

[7] **Yifei Wang***, Qi Zhang*, Yaoyu Guo, and Yisen Wang. Non-negative Contrastive Learning. In *ICLR*. 2024.

[8] **Yifei Wang***, Jizhe Zhang*, and Yisen Wang. Do Generated Data Always Help Contrastive Learning? In *ICLR*. 2024.

[9] **Yifei Wang***, Liangchen Li*, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. Balance, Imbalance, and Rebalance: Understanding Robust Overfitting from a Minimax Game Perspective. In *NeurIPS*. 2023.

[10] **Yifei Wang***, Yuyang Wu*, Zeming Wei, Stefanie Jegelka, and Yisen Wang. A Theoretical Understanding of Self-Correction through In-context Alignment. In *NeurIPS*. 2024. ***Best Paper Award at ICML 2024 ICL Workshop***.

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

[12] Qi Zhang*, **Yifei Wang***, and Yisen Wang. How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders. In *NeurIPS* ***(Spotlight presentation)***. 2022.

[13] Qixun Wang*, **Yifei Wang***, Hong Zhu, and Yisen Wang. Improving Out-of-distribution Robustness by Adversarial Training with Structured Priors. In *NeurIPS* ***(Spotlight presentation)***. 2022.

[14] Zhijian Zhuo*, **Yifei Wang***, Jinwen Ma, and Yisen Wang. Towards a Unified Theoretical Understanding of Non-contrastive Learning via Rank Differential Mechanism. In *ICLR*. 2023.

[15] Rundong Luo*, **Yifei Wang***, and Yisen Wang. Rethinking the Effect of Data Augmentation in Adversarial Contrastive Learning. In *ICLR*. 2023.

[16] Xiaojun Guo*, **Yifei Wang***, Tianqi Du, and Yisen Wang. ContraNorm: A Contrastive Learning Perspective on Oversmoothing and Beyond. In *ICLR*. 2023.

[17] Qi Zhang*, **Yifei Wang***, and Yisen Wang. On the Generalization of Multi-modal Contrastive Learning. In *ICML*. 2023.

[18] Jingyi Cui*, Weiran Huang*, **Yifei Wang***, and Yisen Wang. Rethinking Weak Supervision in Helping Contrastive Representation Learning. In *ICML*. 2023.

[19] Xiaojun Guo*, **Yifei Wang***, Zeming Wei, and Yisen Wang. Architecture Matters: Uncovering Implicit Mechanisms in Graph Contrastive Learning. In *NeurIPS*. 2023.

[20] Qi Zhang*, **Yifei Wang***, and Yisen Wang. Identifiable Contrastive Learning with Automatic Feature Importance Discovery. In *NeurIPS*. 2023.

[21] Ang Li*, **Yifei Wang***, Yiwen Guo, and Yisen Wang. Adversarial Examples Are Not Real Features. In *NeurIPS*. 2023.

[22] Tianqi Du*, **Yifei Wang***, and Yisen Wang. On the Role of Discrete Tokenization in Visual Representation Learning. In *ICLR* ***(Spotlight presentation)***. 2024.

[23] George Ma*, **Yifei Wang***, and Yisen Wang. Laplacian Canonization: A Minimalist Approach to Sign and Basis Invariant Spectral Embedding. In *NeurIPS*. 2023.

[24] George Ma*, **Yifei Wang***, Derek Lim, Stefanie Jegelka, and Yisen Wang. A Canonization Perspective on Invariant and Equivariant Learning. In *NeurIPS*. 2024.

[25] Sharut Gupta*, Chenyu Wang*, **Yifei Wang***, Tommi Jaakkola, and Stefanie Jegelka. In-Context Symmetries: Self-Supervised Learning through Contextual World Models. In *NeurIPS*. 2024.

[26] Lin Li, **Yifei Wang**, Chawin Sitawarin, and Michael W. Spratling. OODRobustBench: A Benchmark and Large-Scale Analysis of Adversarial Robustness Under Distribution Shift. In *ICML*. 2024.

[27] Shiji Xin, **Yifei Wang**, Jingtong Su, and Yisen Wang. On the Connection between Invariant Learning and Adversarial Training for Out-of-Distribution Generalization. In *AAAI* ***(Oral presentation)***. 2023.

[28] Zeming Wei, **Yifei Wang**, and Yisen Wang. Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations. In: *arXiv preprint arXiv:2310.06387* (2023). ***Cited over 150 times and featured in Anthropic's blog***.

[29] Xinyi Wu, Amir Ajorlou, **Yifei Wang** (advise), Stefanie Jegelka, and Ali Jadbabaie. On the Role of Attention Masks and LayerNorm in Transformers. In *NeurIPS*. 2024.

[30] Qi Zhang, Tianqi Du, Haotian Huang, **Yifei Wang** (advise), and Yisen Wang. Look Ahead or Look Around? A Theoretical Comparison Between Autoregressive and Masked Pretraining. In *ICML*. 2024.

[31] Hanqi Yan, Yanzheng Xiang, Guangyi Chen, **Yifei Wang** (advise), Lin Gui, and Yulan He. Encourage or Inhibit Monosemanticity? Revisit Monosemanticity from a Feature Decorrelation Perspective. In *EMNLP*. 2024.

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

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

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

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

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

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