# Kaiwen Hu

https://kaotty.github.io kaiwen_hu@berkeley.edu

## Education

**School of EECS, UC Berkeley** Berkeley, U.S.
*PhD in EECS, advised by Prof. Somayeh Sojoudi* *Aug. 2025 - May 2030 (Expected)*

**School of EECS, Peking University** Beijing, China
*BS in Computer Science, advised by Prof. Yisen Wang* *Sep. 2021 - July 2025*

## Research Interests

- Reasoning in LLMs: Understanding the causes of reasoning failures in large language models and solving them via reliable and robust methods.
- Self-Supervised Learning: Revisiting the existing self-supervised learning paradigms and uncovering their underlying mechanisms theoretically.
- Reinforcement Learning: Developing scalable RL algorithms and applying them to applications like LLMs.

## Publications and Preprints

* indicates equal contributions

- **Understanding the Role of Equivariance in Self-Supervised Learning**
  Yifei Wang*, **Kaiwen Hu***, Sharut Gupta, Ziyu Ye, Yisen Wang, Stefanie Jegelka
  The 38th Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2024.
- **Projection Head is Secretly an Information Bottleneck**
  Zhuo Ouyang*, **Kaiwen Hu***, Qi Zhang, Yifei Wang, Yisen Wang
  The 13th International Conference on Learning Representations (**ICLR**), 2025.
- **Value Gradient Flow: Behavior-Regularized RL without Regularization**
  Haoran Xu*, **Kaiwen Hu***, Somayeh Sojoudi, Amy Zhang
  The 14th International Conference on Learning Representations (**ICLR**), 2026.

## Research Experiences

**Equivariant Self-supervised Learning Theory (Accepted at NeurIPS 2024)** Oct. 2023-May 2024

*Advised by* **Prof. Yisen Wang (School of Artificial Intelligence, Peking University)**

- **Summary:** In this work, we seek to establish a theoretical explanation for the principle of equivariant self-supervised learning (E-SSL) using the information theory. We utilize the explaining-away effect to analyze the mutual information between augmentation and class information, which indicates that E-SSL can indeed encourage the model to extract class-relevant features from data and benefit downstream tasks. We also propose three principles for the design of E-SSL and explain how advanced E-SSL designs echo with our theory.

  **Contributions:**

- Use a toy model to investigate how to maximize the mutual information between the augmentation information and the class information when given the encoder feature.
- Discover that a balanced mixture of augmentation information and class information makes a good feature and that a larger action space is preferable for learning a better feature.
- Verify that acquiring class information can help E-SSL by explicitly injecting and eliminating class information during pretraining.
- Verify that aggressive base transformations can effectively prune shortcuts like learning from colors in E-SSL, encouraging the model to rely more on extracting class information during pretraining.

**Contrastive Learning Theory (Accepted at ICLR 2025)** June 2024-Oct. 2024

*Advised by* **Prof. Yisen Wang (School of Artificial Intelligence, Peking University)**

- **Summary:** In this work, we aim to investigate the role of the projection head in contrastive learning from an information theory perspective. We establish theoretical lower and upper bounds for the downstream performance, suggesting that the projection head should act as an information bottleneck. In addition, we optimize the design of the projection head by means of training and structural regularization, which improves downstream performance.

  **Contributions:**

- Establish the theoretical lower and upper bounds on downstream performance and propose that the key point is to control the mutual information between the encoder and projector features.
- Add the matrix mutual information between the encoder and projector features as a regularization term to the contrastive loss, which increases downstream task accuracy and aligns well with theory.
- Discover that discretizing the projector feature improves downstream performance across different datasets and propose a brief theoretical explanation for this method.

**Behavior-regularized Reinforcement Learning (Accepted at ICLR 2026)** Nov. 2024-Sep. 2025

*Advised by* **Prof. Amy Zhang (Electrical and Computer Engineering Department, UT Austin)**

- **Summary:** In most RL settings, we need a KL divergence term to control the distance between the reference policy and the learned policy, but it inherently limits the support of the learned policy to that of the reference policy. In this work, we introduce Value Gradient Flow (VGF), a particle-based gradient flow method to learn the policy, which does not apply any regularization term. We theoretically prove that VGF has an impilcit regularization by controlling the VGF learning step, and is able to explore beyond the support of the reference policy. VGF demonstrates superior performance on RL benchmarks such as D4RL and OGBench.

  **Contributions:**

- Theoretically prove that VGF implicitly imposes regularization on the learned policy by showing that the Maximum Mean Discrepancy (MMD) has an upper bound that is linear with respect to the VGF learning rate and the VGF step.
- Theoretically prove that VGF can enable the learned policy to explore beyond the support of the reference policy by showing that the $\epsilon$-support of the learned policy is almost surely not a subset of that of the reference policy in both discrete and continuous action spaces.
- Evaluate the performance of VGF on D4RL and OGBench.

**Rethinking the Structure and Mistakes in Reasoning (Ongoing project)** Oct. 2025-Present

*Advised by* **Prof. Sewon Min (Electrical Engineering and Computer Science Department, UC Berkeley)**

- **Summary:** In this work, we propose a fine-grained hierarchical taxonomy of error types and construct a benchmark to evaluate the performance of state-of-the-art LLMs. We also introduce a multi-stage pipeline that prompts the LLM to first classify the question to a certain domain, then perform iterative verification on each step, and finally classify the error to a specific domain-related type if an error is located. Using DeepSeek-V3.2 as our annotator model, we find that the overall error distribution vary with model families and observe that DeepSeek-R1 demonstrates superior in-place-recovery and post-recovery rates over other models.

  **Contributions:**

- Establish a two-level error taxonomy of reasoning traces and manually construct a benchmark to evaluate the annotation capability of state-of-the-art LLMs.
- Propose a multi-stage pipeline that first classify the problem into a domain, then verify the correctness of each step with previous steps as reference, and finally classify the errors.
- Use DeepSeek-V3.2 as the annotator model and find that "Misidentification & Hallucination" and "Unjustified Geometric Property & Relationship" are the most prevalent errors.

## Skills

- Programming Languages: C, C++, Python (Pytorch).
- Tools: VS code, Github, Latex, Notions, Markdown.

## Honors and Awards

| | |
|---|---|
| Third Prize in the Peking University Programming Contest | 2023 |
| Peking University Excellent Study Award | 2022 |
| First Prize in the National High School Mathematics Competition (Shanghai Provincial) | 2019, 2020 |