Consider a test of causal understanding: if we remove an object from an image, the shadows should change accordingly. Modern generative models can synthesize photorealistic images yet consistently fail at such reasoning tasks. This reveals a fundamental limitation – today's models learn correlations from data without internalizing the causal processes that generate it, limiting out-of-distribution generalization and real-world robustness of AI. My research addresses this by developing **principled representation learning for generative models** to recover causally-grounded and robust latent factors from abundant raw sensory data.

During PhD, I aim to explore design principles for generative models where causal understanding of the world emerges naturally in learned representations. As natural data contains rich regularities (low-dimensionality, spatiotemporal smoothness, etc.), these structures can guide the learning process toward robust rather than spurious representations. I believe this is a viable path towards trustworthy AI systems with robustness and compositional generalizability, without massive human supervision or computational power. This has concrete applications: from robotic planning under uncertainty to generating novel hypotheses for scientific discovery, where causal understanding matters more than pattern matching.

Past Research

Instead of focusing solely on engineering aspects of AI systems, my research revisits fundamental modeling assumptions and develops practical solutions from first principles, drawing ideas from various fields such as signal processing, classical AI, or deep learning. This approach produced **4 first-author publications** at top venues in **machine learning (NeurIPS)** and **signal processing (ICASSP, SAM)**, spanning **representation learning** and **graph learning**.

1) Identifiable Representation Learning

With **Prof. Xiao Fu** at **Oregon State University**, I studied a fundamental question: can we learn the underlying causal factors behind high-dimensional observations without supervision? Consider medical imaging, where factors like disease severity or and imaging artifacts are entangled in each scan. Disentangling these factors would enable counterfactual reasoning (a scan with vs. without disease) and robust diagnostics. However, this problem is ill-posed: without constraints, many spurious representations exist that fit the data.

Existing methods impose strong assumptions: some require independent factors given weak supervision, while others demand influence sparsity (each factor affects only a few observed features). These fail in many scenarios: auxiliary data is not always available, and strict sparse influence cannot describe global-effect factors like lighting. My **NeurIPS 2025** paper introduces *diverse influence component analysis* (DICA), which achieves identifiability through a novel principle – real-world factors influence observed data in geometrically diverse ways, bypassing previous restrictive assumptions. Via Jacobian-based regularization, DICA enforces a structured latent space inside generative model, allowing provable extraction of causal representations behind data.

The core insight emerged after months of exploring connections between generative modeling and signal processing. I realized that if two factors control distinct aspects of data generation (say, object position versus lighting), their partial derivative vectors should span quite different directions. Formalizing this intuition into rigorous identifiability proofs required connecting DICA to matrix factorization, which I discovered through extensive literature study, from differential geometry to source separation. The result is an elegant learning criterion: constraining generator's Jacobian such that each factor’s influence on data manifolds is geometrically diverse. Implementing DICA at scale is difficult: optimizing Jacobian determinants for deep neural networks is expensive and unstable. To make the objective tractable, I developed a stochastic approximator using matrix sketching techniques, drawing on my advanced coursework in optimization. With this scalable implementation, we demonstrated DICA's effectiveness on two challenging applications: inferring transcription

factor activities in single-cell genomics and disentangling visual concepts from images without labels. This work exemplifies my approach of revisiting classical ideas and scaling them to deep learning.

2) Robust Graph Learning

The insight that structure in data can fend off ill-posedness and missing information crystallized in my earlier work with **Prof. Hoi-To Wai** at **Chinese University of Hong Kong**. Real-world graph datasets inevitably contain missing nodes due to incomplete data collection or privacy constraints. My work leverages *graph signal smoothness*, inherent in many real networks, to handle partial observation.

At **SAM 2024** (**Best Student Paper Award**), I developed a method to estimate signal smoothness without ground-truth topology, under partial observation. After deep experiments, I discovered that smoothness manifests in spectral properties of feature covariance, even when nodes are missing. Hence, I proposed a simple practical algorithm: first estimate smoothness from features alone, then use it to filter corrupted training samples for downstream tasks such as community detection. My follow-up work (**GSP Workshop 2025**, sent to **ICASSP 2026**) provides the first stability analysis of graph topology learning under partial observation. Interestingly, we showed that simple hidden-nodes-agnostic methods can match complicated models when signals are sufficiently smooth. This connects to my broader research: identifying minimal structures to make problems tractable.

Future Research

Building on my research in identifiable and robust representation learning, my future work aims to bridge the gap between causal understanding and generative modeling. A critical bottleneck in current decision-making systems is their reliance on spurious correlations to maximize rewards, rather than internalizing the causal dynamics of the environment. To address this, I propose two directions. First, I aim to build generative world models for robust planning. By constraining these models with physical regularities, I can build reliable internal simulators without massive labels, enabling agents to plan effectively under uncertainties. Second, I investigate new architectures that align shared representations across unpaired modalities. While current approaches depend on paired data, I aim to formulate alignment objectives that leverage the world's shared causal structure to learn from abundant unpaired data. These directions allow me to translate my theoretical background into tangible, physically-grounded AI.