Beyond Features to Mechanisms: Foundations of Identifiable Representation Learning

Modern AI systems excel at modeling and generating high-dimensional data by learning expressive representations. Yet these systems largely remain black boxes where the learned features lack a correspondence to the **true underlying mechanisms** we care about, *e.g.,* physical laws governing a system, causal relations driving an observed phenomenon, or factors of variation we wish to control. This opacity is a fundamental obstacle to the reliability of AI models. A core culprit is *non-identifiability*: under current paradigms, wildly different but equally plausible representations can explain the same observations, obscuring what the model has learned and whether its behavior will remain stable when conditions change. This underspecification leads to brittleness under even minor distribution shifts [1], impeding AI's use in safety-critical scientific and engineering domains.

My research asks a central question: *What is the right foundation for learning representations we can trust?* I argue for a shift from merely predictive features to representations that faithfully model the data-generating process, yielding two desiderata for an ideal representation: (i) **identifiable**—the object of learning is well-defined and unique (up to simple equivalences, *e.g.,* scaling factors); and (ii) **causal**—it captures cause-effect relationships, *e.g.,* enabling answers to "what if" questions. When these conditions hold, the representation is *explainable by construction*, with components aligned to real mechanisms rather than opaque correlations.

My research vision is to establish learning frameworks that deliver identifiable, causal, and explainable representations, grounded in rigorous theory and validated on real-world problems.

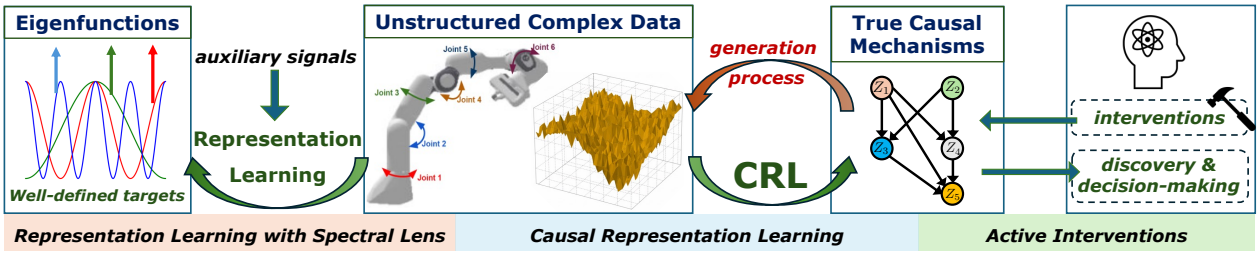

Fig. 1: Overview of my research toward identifiable, causal representations and their use in decision-making.

1. Causal Representation Learning (CRL) [2–7]: At the center is **CRL**, which models high-dimensional observations as generated from interacting latent causal mechanisms. This view extends classic methods such as independent component analysis (ICA) and aligns with the scientific goal of uncovering the mechanisms behind observations. I pioneered *interventional CRL* to provably recover these mechanisms by leveraging multiple *environments* [2–6], where an "intervention" is simply a change (passively observed or actively applied) in the mechanisms across environments. My work provides theory-grounded, scalable algorithms for real-world problems, for instance, estimating robot pose from raw images without labels [7]. These foundations drive my plans to: (i) develop a **unified CRL theory** integrating diverse knowledge sources (multi-environment, time series, multi-view data), and (ii) tackle **high-impact applications** (*e.g.,* robot perception, scientific imaging) as CRL tasks, where recovering true mechanisms enables transfer across settings.

2. Causal Learning with Active Interventions [8–14]: My work addresses how to learn and decide efficiently via active interventions—a common task in domains from clinical care to scientific discovery. With causal variables in hand (*e.g.,* from CRL), I show that exploiting causal structure yields exponential gains over non-causal baselines in sequential intervention design. I also design adaptive algorithms to *discover* causal structures in challenging regimes (*e.g.,* heterogeneous populations) where passive data are insufficient.

3. Representation Learning through a Spectral Lens [15, 16]: Beyond explicitly causal settings, my work provides an identifiability theory for representation learning at large. We formalize learning objectives as pairing each input with a *context* (*e.g.,* label, masked token, or augmented view), and show that learned representations align with the most stable components of this relationship, formally, the top eigenfunctions of an induced kernel. My work sharpens this spectral perspective into theoretical guarantees (*e.g.,* optimality) and empirical evidence.

Causal Representation Learning (CRL)

The central theoretical challenge in CRL is *identifiability*—uniquely recovering the true causal variables—which is famously impossible when using data from a single, static distribution, *i.e.,* lacking sufficient statistical diversity [17]. I show that identifiability becomes possible by using data from multiple related environments, as sparse changes across environments (*i.e.,* interventions) create the needed diversity in the observed data.

Theoretical Foundations: A Score-based Framework [2–6]. To harness the additional knowledge induced by the interventions, I developed **score-based CRL**. The key idea is that latent score functions (gradients of log-densities) encode intervention effects as sparse variations across environments. By linking score functions of observed data (which can be efficiently estimated even in high-dimensional settings) to latent scores, I obtain tractable learning objectives and algorithms. In short, my algorithms search for representations whose inter-environment differences are maximally sparse, aligning them with the true causal mechanisms.

The theoretical foundations of CRL have two main axes: (i) model complexity, *i.e.,* how expressive the true causal mechanisms and the mapping from latent to observed data are; and (ii) data richness, *i.e.,* what diversity of environments and interventions is necessary and sufficient for identifiability. This perspective clarifies which structural assumptions and data resources are needed in different regimes.

Using this score-based framework, I established strong theoretical guarantees across this spectrum of settings. I show fully nonparametric identifiability results (for arbitrarily complex latent mechanisms and mapping functions) and provide **the first provably correct algorithm** for this setting [2, 3]. Under *linear observation maps*–a common assumption in related literature, such as ICA and factor analysis—I show that identifiability is possible with fewer interventions than in the general case, and I provide a practical learning algorithm that achieves this result [2]. I further extend this framework to handle challenging and realistic *multi-target* interventions [4], where an unknown subset of mechanisms can change, and provide provably correct algorithms for this regime. My work also advances the statistical understanding of CRL by providing the first *sample-complexity* guarantees [5], pinpointing accurate score estimation as the main statistical bottleneck.

Application: Robotics [7]: Score-based CRL offers a new lens for solving real-world inverse problems. In [7], we address robot pose estimation, *i.e.,* recovering a robot's joint angles from raw camera images. From a multi-environment perspective, each change in the robot's actuator commands induces a new environment that alters only a small subset of joint-angle mechanisms, providing the data diversity needed for score-based CRL. Our method then recovers controllable joint angles directly from images. As a strong validation of identifiable representation learning, this theory-driven, label-free approach achieves performance competitive with state-of-the-art supervised methods [18].

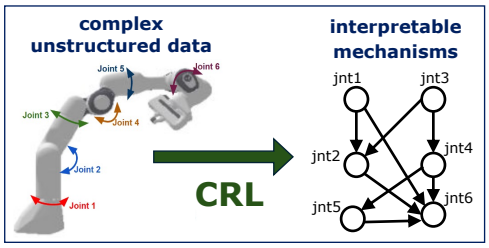

Fig. 2: CRL for Robot Pose Estimation: CRL recovers the key latent variables of the pose (joint angles) from unlabeled raw images.

Causal Learning with Active Interventions

Complementing CRL, this thrust handles settings where causal variables (or representations) are available—either recovered via CRL or observed directly. Many practical goals then require answering causal queries in this space (*e.g.,* which treatment maximizes patient recovery, or whether two factors are causally related). Since observational data alone cannot always resolve such questions, interventions are essential, and the challenge is to learn with minimal interventions (*e.g.,* few types of interventions, few trials, limited complexity). In medicine, for example, it is crucial to find the best treatment with as few clinical trials as possible, where each treatment may involve multiple interacting variables. I develop data-adaptive strategies that efficiently select interventions for such problems. I apply these strategies to the following two broad settings.

Sequential Intervention Design [8–10]: How should we *choose* interventions to efficiently maximize a reward (*e.g.,* patient recovery)? By exploiting given or learned causal knowledge, my algorithms achieve *exponential* improvements over non-causal baselines [8]. Instead of exhaustively testing all possible interventions (combinatorially many), the causal approach allows us to focus on a small set of options that are most informative about the optimal one. Crucially, these gains require no prior knowledge of intervention distributions, which was a key limitation of prior work. We also address the practical challenge of *robustness*. Real-world systems can fluctuate over time, *e.g.,* due to measurement error. To handle this, we developed the first *robust* algorithms for this intervention design problem [9, 10], which remain near-optimal under such temporal changes, with proven performance–robustness trade-offs.

Causal Discovery [11–14]: Real data are often heterogeneous, for example, due to patient subpopulations. This heterogeneity can be modeled as mixtures of causal graphs. Such mixing, however, induces spurious associations that observational data cannot separate from true causal effects. I characterized *when* mixtures of causal graphs create such ambiguities [11], and designed algorithms that adaptively choose interventions to resolve them and learn the *true* underlying cause-effect relationships [12]. I also addressed a common practical challenge in targeted causal discovery: given two environments, can we quickly identify which components have changed? This is especially important for large-scale systems, *e.g.,* localizing root causes of failures in distributed systems. I designed consistent, hierarchical algorithms that achieve at least an order-of-magnitude speed-up over prior methods and provide the first sample-complexity guarantees for this task [13, 14].

Representation Learning through a Spectral Lens

Modern pretraining excels empirically, yet we still lack a systematic account of the learned representations: *what* they capture, *why* they transfer across certain tasks, and *when* they are optimal. Complementing CRL, I recast common representation learning paradigms from an identifiability perspective to answer these questions and propose novel methods to increase the efficiency and extent of identifiability.

Learning from Contexts [15]: We develop *contexture theory* to formalize the implicit mechanisms behind modern representation learning. At a high level, various paradigms can be understood through a single lens: they all rely on learning from *input–context* pairs, where context can take different forms, *e.g.,* labels for supervised learning, masked/corrupted views for self-supervised learning, and local neighborhoods for manifold learning. Despite their differences, our central result is that they all share a common target, that is, **approximating the eigenfunctions of an input–context kernel**. Building on this result, we introduce "task-context compatibility" to specify when a learned representation is optimal for a downstream task. The broader significance of this unified view is twofold: (i) it provides a clear explanation for why representation alignment occurs across different models and data modalities, and (ii) it implies that once models are large enough to approximate the target eigenfunctions, further scaling has diminishing returns, shifting the focus for progress toward designing better, more informative contexts rather than merely scaling up.

Ordered and Identifiable Representations [16]: Our results in [15] expose a key limitation of the common paradigms: standard training objectives typically recover an *entangled eigensubspace* instead of individual, ordered eigenfunctions. Because eigenfunctions differ in importance (due to different eigenvalues) and smoothness, this entanglement mixes their order, blurs high- and low-importance components, and often creates optimization difficulties. In [16], I propose a general framework for learning **ordered eigenfunctions** as identifiable, disentangled representations and provide scalable algorithms that subsume common contrastive and non-contrastive objectives. Due to this ordering property, representations can be truncated on demand to match possible resource constraints. This yields *adaptive-dimensional* representations, enabling controllable accuracy–efficiency trade-offs across settings, from resource-constrained edge devices to high-performance systems.

Future Directions

My work thus far lays the foundation for a long-term research program in identifiable representation learning. I will broaden these theoretical foundations to solve high-impact problems and build robust AI systems.

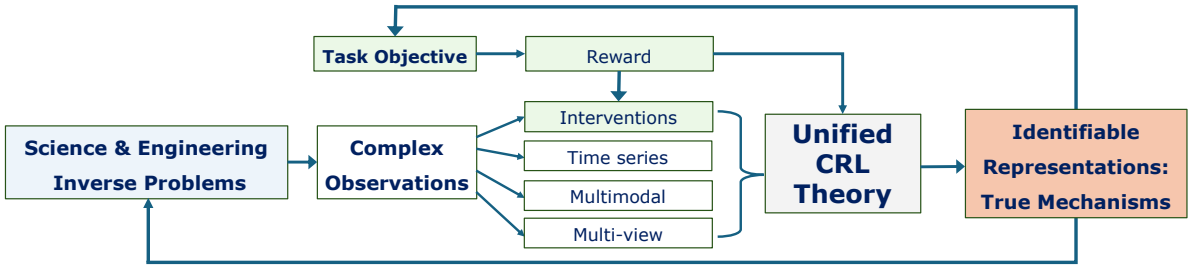

Fig. 3: Future Plans: (i) a unified CRL theory to integrate diverse data sources, (ii) applying this theory to solve high-impact problems, and (iii) representation-guided agents to synergize representation learning with downstream decision-making.

Unified Theory for CRL: My research has pioneered interventional CRL by exploiting variability across *environments* created by interventions. In many applications (*e.g.,* robotics), data also arrive from other sources, such as time series (*e.g.,* videos), multi-view data, and multimodal sensors. The field currently develops isolated theoretical and practical methods for each source, and this fragmentation is a critical bottleneck limiting the potential of CRL. **I will develop a unified CRL theory** that treats these sources as *complementary signals* for learning the underlying true mechanisms. The goal is to combine, for example, multi-environment variation with temporal smoothness to relax the theoretical requirements each source imposes when used alone. My score-based approach [2–4] is an ideal foundation for achieving this goal. Score functions provide a shared object that can model interventional, temporal, and multimodal dependencies, yielding a single mathematical framework to leverage these diverse sources together. The outcome will be a unified CRL toolkit capable of learning identifiable, robust causal representations that transfer across environments and support causal reasoning.

CRL for Scientific Inverse Problems: Many scientific and engineering challenges (*e.g.,* protein imaging, materials discovery, earthquake monitoring) are high-dimensional inverse problems: inferring physical factors (z) from complex observations (x). Standard deep learning often yields brittle, ill-posed mappings ($x \to z$) that generalize poorly. Even powerful generative approaches (*e.g.,* diffusion models) are primarily used to find plausible solutions—not the true mechanisms. **I propose to reframe these challenges as CRL problems.** This is a compelling fit, as inverse problems aim to infer the latent variables/parameters of a forward model from indirect observations. My foundational CRL work is well-suited to exploit this connection. The diverse data that scientists and engineers already collect (*e.g.,* via different sensors or experimental probes) provide the statistical diversity CRL requires for recovery guarantees. Concretely, within my unified CRL framework I will enforce mechanism stability across environments, yielding representations that are (i) identifiable up to natural symmetries, (ii) robust across sites and protocols, and (iii) actionable for causal reasoning and optimal experiment design—ultimately shifting from brittle features to stable and controllable mechanisms.

Causal Representation-Guided Agents: So far, I have discussed learning causal, identifiable representations from *given* data. But how should a learner or agent decide what data to gather, especially in domains like robotics where learning occurs through interactions toward an objective? I will build *Causal Representation-Guided Agents* that actively choose interventions (*e.g.,* manipulating an object) to jointly learn a causal world model and act to obtain rewards. This approach closes the loop between representation and action: interactions generate diverse data (*e.g.,* multi-environment and/or time series), the unified CRL theory converts them into mechanism-aligned representations, and sequential intervention design supplies efficient policies [8, 9]. I plan to ground this direction in robotics, extending my recent CRL validation testbed [7]. Ultimately, I aim to build a framework where CRL and policy learning co-evolve, yielding task-sufficient, controllable representations tailored to the agent's objectives.

References

[1] Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

[2] **Burak Varıcı**, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

[3] **Burak Varıcı**, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *Proc. International Conference on Artificial Intelligence and Statistics*, Valencia, Spain, May 2024.

[4] **Burak Varıcı**, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

[5] Emre Acartürk, **Burak Varıcı**, Karthikeyan Shanmugam, and Ali Tajer. Sample complexity of interventional causal representation learning. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

[6] **Burak Varıcı**, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv:2301.08230*, 2023.

[7] Pranamya Kulkarni, Puranjay Datta, **Burak Varıcı**, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. ROPES: Robotic pose estimation via score-based causal representation learning. *arXiv:2510.20884*, 2025.

[8] **Burak Varıcı**, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *Journal of Machine Learning Research*, 24(297):1–59, 2023.

[9] Zirui Yan, Arpan Mukherjee, **Burak Varıcı**, and Ali Tajer. Robust causal bandits for linear models. *IEEE Journal on Selected Areas in Information Theory*, 5:78–93, 2024.

[10] Zirui Yan, Arpan Mukherjee, **Burak Varıcı**, and Ali Tajer. Improved bound for robust causal bandits with linear models. In *Proc. International Symposium on Information Theory*, Athens, Greece, 2024.

[11] **Burak Varıcı**, Dmitriy Katz-Rogozhnikov, Dennis Wei, Prasanna Sattigeri, and Ali Tajer. Separability analysis for causal discovery in mixture of DAGs. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[12] **Burak Varıcı**, Dmitriy A Katz, Dennis Wei, Prasanna Sattigeri, and Ali Tajer. Interventional causal discovery in a mixture of DAGs. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

[13] **Burak Varıcı**, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Scalable intervention target estimation in linear models. In *Proc. Advances in Neural Information Processing Systems*, December 2021.

[14] **Burak Varıcı**, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Intervention target estimation in the presence of latent variables. In *Proc. Conference on Uncertainty in Artificial Intelligence*, Eindhoven, Netherlands, August 2022.

[15] Runtian Zhai, Kai Yang, **Burak Varıcı**, Che-Ping Tsai, J. Zico Kolter, and Pradeep Ravikumar. Contextures: Representations from contexts. In *Proc. International Conference on Machine Learning*, Vancouver, Canada, July 2025.

[16] **Burak Varıcı**, Che-Ping Tsai, Ritabrata Ray, Nicholas M. Boffi, and Pradeep Ravikumar. Eigenfunction extraction for ordered representation learning. *arXiv:2510.24672*, 2025.

[17] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. International Conference on Machine Learning*, Long Beach, CA, June 2019.

[18] Raktim Gautam Goswami, Prashanth Krishnamurthy, Yann LeCun, and Farshad Khorrami. RoboPEPP: Vision-based robot pose and joint angle estimation through embedding predictive pre-training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, June 2025.