Relation-First Modeling Paradigm for Causal Representation Learning toward the Development of AGI

Anonymous authors
Paper under double-blind review

Abstract

The traditional i.i.d.-based learning paradigm faces inherent challenges in addressing causal relationships, which has become increasingly evident with the rise of applications in causal representation learning. Our understanding of causality naturally requires a perspective as the creator rather than observer, as the "what...if" questions only hold within the possible world we conceive. The traditional perspective limits capturing dynamic causal outcomes and leads to compensatory efforts such as the reliance on hidden confounders. This paper lays the groundwork for the new perspective, which enables the *relation-first* modeling paradigm for causality. Also, it introduces the Relation-Indexed Representation Learning (RIRL) as a practical implementation, supported by experiments that validate its efficacy.

1 Introduction

The concept of Artificial General Intelligence (AGI) has prompted extensive discussions over the years yet remains hypothetical, without a practical definition in the context of computer engineering. The pivotal question lies in whether human-like "understanding", especially causal reasoning, can be implemented using formalized languages in computer systems Newell (2007); Pavlick (2023); Marcus (2020). From an epistemological standpoint, abstract entities (i.e., perceptions, beliefs, desires, etc.) are prevalent and integral to human intelligence. However, in the symbol-grounded modeling processes, variables are typically assigned as observables, representing tangible objects to ensure their values have clear meaning.

Epistemological thinking is often anchored in objective entities, seeking an irreducible "independent reality" Eberhardt & Lee (2022). This approach necessitates a metaphysical commitment to constructing knowledge by assuming the unproven prior existence of the "essence of things", fundamentally driven by our desire for certainty. Unlike physical science, which is concerned with deciphering natural laws, technology focuses on devising effective methods for problem-solving, aiming for the optimal functional value between the nature of things and human needs. This paper advocates for a shift in perspective when considering technological or engineering issues related to AI or AGI, moving from traditional epistemologies to that of the creator. That is, our fundamental thinking should move from "truth and reality" to "creation and possibility".

In some respects, both classical statistics and modern machine learnings traditionally rely on epistemology and follow an "object-first" modeling paradigm, as illustrated by the practice of assigning pre-specified, unchanging values to variables regardless of the model chosen. In short, individual *objects* (i.e., variables and outcomes) are defined a priori before considering the *relations* (i.e., model functions) between them by assuming that what we observe precisely represents the "objective truth" as we understand it. This approach, however, poses a fundamental dilemma when dealing with causal relationship models.

Specifically, "causality" suggests a range of possible worlds, encompassing all potential futures, whereas "observations" identify the single possibility that has actualized into history with 100% certainty. Hence, addressing causal questions requires us to adopt the perspective of the "creator" (rather than the "observer"), to expand the objects of our consciousness from given entities (i.e., the observational world) to include possible worlds, where values are assigned "as supposed to be", that is, *as dictated by the relationship*.

Admittedly, causal inference and related machine learning methods have made significant contributions to knowledge developments in various fields Wood (2015); Vuković (2022); Ombadi et al. (2020). However, the inherent misalignment between the “object-first” modeling principle and our instinctive “relation-first” causal understanding has been increasingly accentuated by the application of AI techniques, i.e., the neural network-based methods. Particularly, integrating causal DAGs (Directed Acyclic Graphs), which represent established knowledge, into network architectures Marwala (2015); Lachapelle et al. (2019) is a logical approach to efficiently modeling causations with complex structures. However, surprisingly, this integration has not yet achieved general success Luo et al. (2020); Ma et al. (2018).

As Scholkopf Schölkopf et al. (2021) points out, it is commonly presumed that “the causal variables are given”. In response, they introduce the concept of “causal representation” to actively construct variable values as causally dictated, replacing the passively assumed observational values. However, the practical framework for modeling causality, especially in contrast to mere correlations, remains underexplored. Moreover, this shift in perspective suggests that we are not just dealing with “a new method” but rather a new learning paradigm, necessitating in-depth philosophical discussions. Also, the potential transformative implications of this “relation-first” paradigm for AI development warrant careful consideration.

This paper will thoroughly explore the “relation-first” paradigm in Section 2, and introduce a complete framework for causality modeling by adopting the “creator’s” perspective in Section 3. In Section 4, we will propose the *Relation-Indexed Representation Learning* (RIRL) method as the initial implementation of this new paradigm, along with extensive experiments to validate RIRL’s effectiveness in Section 5.

2 Relation-First Paradigm

The “do-calculus” format in causal inference Pearl (2012); Huang (2012) is widely used to differentiate the effects from “observational” data X, and “interventional” data $do(X)$ Hoel et al. (2013); Eberhardt & Lee (2022). Specifically, $do(X = x)$ represents an intervention (or action) where the variable X is set to a specific value x, distinct from merely observing X taking the value x. However, given the causation represented by $X \to Y$, why doesn’t $do(Y = y)$ appear as the action of another variable Y?

Particularly, compared to the ***static*** state X, $do(X)$ incorporates its “temporal aspect” to represent the process of “becoming X” as a ***dynamic*** object. This concept is analogous to $do(Y)$, with the naturally understood relationship $do(X) \to do(Y)$ being formed. Take, for instance, the statement “The storm lasting for a week causes downstream villages to be drowned by the flood”, if $do(X)$ represents the storm lasting a week, then $do(Y)$ could be considered as the ensuing disaster of the villages drowning.

The difficulty of involving $do(Y)$ stems from the empirical difficulties when building models. We can naturally envision $do(Y)$ before it becomes true, in a “possible world” that we freely *create* in our consciousness, with its timing or states varying. For example, the disaster denoted by $do(Y)$ could happen earlier or later, with different levels of severity, depending on various conditions. This is in contrast to the observational world of $do(X)$, where only a single “actual timing” exists, allowing us to focus on modeling observed states at specific timestamps, like X_t, without the need to account for “varying timing”.

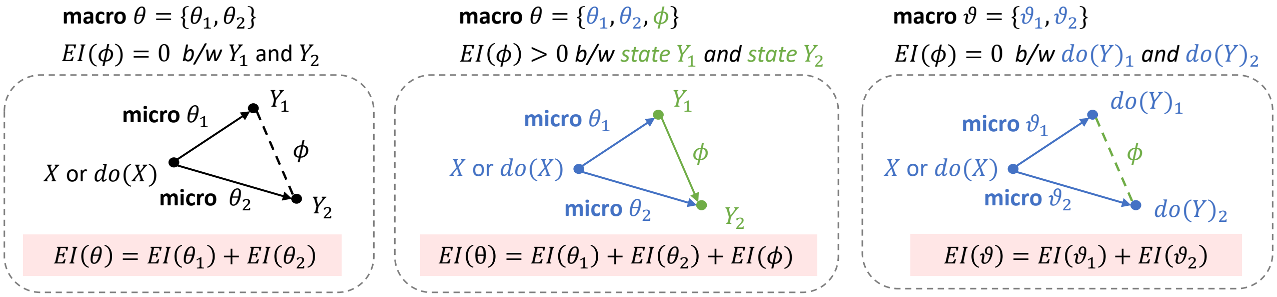
However, it does not mean the distributions over timing are insignificant for modeling. On the contrary, the inclusion of $do(X)$ highlights its necessity. Particularly, RNNs (Recurrent neural networks) can autonomously identify the significant dynamics to form the “cause”, effectively achieving $do(X) \to Y$, without the need to manually pre-specify “identifiable $do(X)$” as statistical causal inference often requires Pearl (2012) (e.g., determining the duration of a storm necessary to qualify as a significant $do(X)$ for different watersheds).

In RNNs, $do(X)$ is optimized in the form of latent space features, by being related to the outcome Y. These features were initialized to represent the sequence $X^t = [X_1, \ldots, X_t]$ with determined timestamp t. But, as representations rather than mere observational variables, they allow the t-dimension to be incorporated into computations to evaluate the t value’s significance. The capability of RNNs to effectively realize $do(X) \to Y$ has led to their growing popularity in relationship modeling Xu et al. (2020). However, the question arises: Can the same approach be used to autonomously extract $do(Y)$ without pre-specifying its timing?

Since the technique has emerged, it may be time for us to see the “possible world” in computations. This requires a shift in perspective from being the “observer” to becoming a “creator,” essential for engaging in causal modeling within a “relation-first” paradigm. This section will delve into its philosophical underpinnings, mathematically redefine “causality”, and explore its broader implications.

2.1 Philosophical Foundation

The Causal Emergence theory Hoel et al. (2013); Hoel (2017) represents a significant advancement in the philosophical understanding of causality. It suggests that while causality is often observed in micro-causal components only (e.g., in the single relationship $X \rightarrow Y$), the macro-causal system (e.g., a multi-variable causal system comprising $\{X, Y, Z\}$) can exhibit more informative interactions than just the sum of its components (e.g., the system's behavior cannot be fully explained by simply combining $X \rightarrow Y$, $X \rightarrow Z$, and $Y \rightarrow Z$). The macro- and micro-causality originates from economics, while in modeling contexts, their scales can be flexibly determined based on practical questions. For example, if $X \rightarrow Y$ is viewed as macro-causal, with Y_1 and Y_2 representing two complementary parts of Y such that $Y = Y_1 + Y_2$, it can be divided into two micro-causal components $X \rightarrow Y_1$ and $X \rightarrow Y_2$, as illustrated in Figure 1.


macro $\theta = \{\theta_1, \theta_2\}$
$EI(\phi) = 0$ b/w Y_1 and Y_2
micro θ_1
Y_1
ϕ
X or $do(X)$
micro θ_2
Y_2
$EI(\theta) = EI(\theta_1) + EI(\theta_2)$
macro $\theta = \{\theta_1, \theta_2, \phi\}$
$EI(\phi) > 0$ b/w state Y_1 and state Y_2
micro θ_1
Y_1
ϕ
X or $do(X)$
micro θ_2
Y_2
$EI(\theta) = EI(\theta_1) + EI(\theta_2) + EI(\phi)$
macro $\vartheta = \{\vartheta_1, \vartheta_2\}$
$EI(\phi) = 0$ b/w $do(Y)_1$ and $do(Y)_2$
micro ϑ_1
$do(Y)_1$
ϕ
X or $do(X)$
micro ϑ_2
$do(Y)_2$
$EI(\vartheta) = EI(\vartheta_1) + EI(\vartheta_2)$

(a) *Correlation with static outcome* $X/do(X) \overset{\theta}{\rightarrow} Y$

(b) *Causation with dynamic outcome but described by* $X/do(X) \overset{\theta}{\rightarrow} Y$*, leading to Confounding phenomenon*

(c) *Causation with dynamic outcome* $X/do(X) \overset{\vartheta}{\rightarrow} do(Y)$

Figure 1: Causal Emergence $EI(\phi) > 0$ stems from overlooking the potential existence of $do(Y)$.

In this context, the concept of Effective Information (EI), represented as $EI(X \rightarrow Y)$, is employed to measure the information generated by the system during the transition from the state of X to the state of Y Tononi & Sporns (2003); Hoel et al. (2013). Moreover, the minimum EI that can be exchanged between Y_1 and Y_2 is represented by an abstract variable ϕ, signifying the ability to integrate this crucial information Tononi & Sporns (2003). Regardless of mere statistical dependence between state components Y_1 and Y_2, which is depicted by the dashed line with $EI(\phi) = 0$ in Figure 1(a), the causal emergence highlights that "their informative interaction with $EI(\phi) > 0$ may not be fully encapsulated by $EI(X \rightarrow Y)$" as in Figure 1(b).

It can be explained by the information loss when reducing a *dynamic* outcome $do(Y)$ to be a *static* Y. For ease of understanding, let's consider the reduction from $do(X) \rightarrow do(Y)$ to $X \rightarrow Y$, which can be likened to attributing the precipitation on a specific date (i.e., the X_t value), as the sole cause for the likelihood of a village being flooded 7 days later (i.e., the Y_{t+7} value), regardless of what happened on the other days. From a computational standpoint, given observational variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$, this reduction implies that the interaction between the $\mathbb{R}^{n+1}$ and $\mathbb{R}^{m+1}$ spaces must be simplistically represented between $\mathbb{R}^n$ and $\mathbb{R}^m$.

Still, achieving $do(X) \rightarrow Y$ is feasible by manually identifying $do(X)$ from observed data (e.g., specifying the duration dates for each qualified storm). But for $Y_1 \rightarrow Y_2$, which implies an interaction in a "possible world" with "possible timings," pre-specifying may prove impractical. Suppose Y_1 represents the impact of flood-prevention operations, and Y_2 signifies the water level on any subsequent day "without" these operations. The actual result $do(Y)$ can be dynamically described as "the flood crest originally expected on the 7th day has been mitigated over the following days, preventing a disaster". However, it would be challenging to precisely specify a particular day's water rising for Y_2 "if without" Y_1.

As Hoel emphasizes Hoel (2017), applying information theory to causality enables us to create "nonexistent" or "counterfactual" statements. Indeed, the concept of "information" is intrinsically linked to ***relation*** and independent of the ***objects*** of entities. For example, let's use an abstract variable θ to represent the EI of transitioning from X_t to Y_{t+7} (similar to how ϕ has been defined). Suppose $\theta =$ "flooding", and $EI(\theta) =$ "what a flooding may imply", we can then easily conceptualize $do(X) =$ "continuous storm" as its cause, and $do(Y) =$ "disastrous water rise" as the result in consciousness, without needing specific observations of precipitation value X_t and water level Y_{t+7}. In other words, our comprehension is in a "relation-first" manner, unlike the "object-first" approach typically employed in our modeling processes.

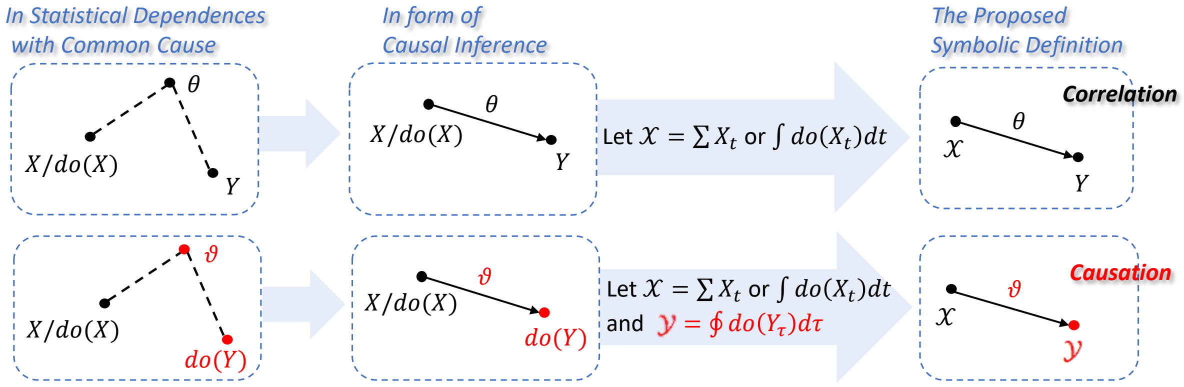
Significantly, the "possible world" we seek to predict is inherently ***created*** by our innate "relation-first" thinking, where *timing* is the essential dimension. Without "possible timing," any "possible observation" would lose its meaning. For example, one might use the model $Y_{t+7} = f(X^t)$ for disaster prediction. However, the true purpose is not "knowing the exact water level on the 7th day" but understanding "how the disaster might unfold; if not on the 7th day, then what about the 8th, 9th, and so on?" Today, advanced techniques capable of addressing this question have emerged, especially with the success of RNNs in computing ***timing distributions*** through latent space representations. Consequently, it is time to reevaluate the conventional definitions and learning paradigm of causality, which were originally established on an "object-first" basis.

Indeed, there has been ongoing progress in mathematically representing the concept of *timing distribution* as aligned with our understanding, but a clear and explicit definition remains elusive. Largely due to being entrenched in epistemology, the timestamp t can only be allowed as a priorly determined "constant" in models rather than a "variable" with computable values. This paper seeks to address this issue fundamentally.

The Markov process established the sequence of independent states over time, denoted as $X^t = (X_1, \dots, X_t)$, like captured through a series of snapshots. Information-theoretic measurements of causality (e.g., directed information Massey et al. (1990), transfer entropy Schreiber (2000)) linguistically emphasized the distinction between perceiving X^t as "a sequence of *static* states" versus holistically as "a *dynamic* transition process". The introduction of do-calculus Pearl (2012) marks a significant advancement, with the notation $do(X_t)$ explicitly treating the action of "becoming X_t" as a *dynamic unit*. However, its nature in differential calculus necessitates each "identifiable" $do(X)$ be manually specified. On the other hand, the notation $do(Y_t)$, which implies a "possibly valued" t, lacks a foundation for declaration. It can only be described in terms of the distribution of future states "informatively constrained" by current ones Hoel et al. (2013); Hoel (2017), but still risks being critiqued for lacking "metaphysical commitments" Eberhardt & Lee (2022).

This paper doesn't intend to engage in metaphysical debates; rather, it aims to highlight that for technological inquiries, shifting from an epistemological to a "creator's" perspective can produce models that align with our instinctive understanding, and also dramatically simplify the notations. These aspects are especially crucial in the context of causal-reasoning AI or AGI. For purely philosophical discussions, readers are encouraged to explore the "creationology" theory by Mr.Zhao Tingyang.

2.2 Mathematical Definition of Relation



In Statistical Dependences with Common Cause
In form of Causal Inference
The Proposed Symbolic Definition
θ
$X/do(X)$
Y
θ
$X/do(X)$
Y
Let $\mathcal{X} = \sum X_t$ or $\int do(X_t)dt$
Correlation
θ
$\mathcal{X}$
Y
ϑ
$X/do(X)$
$do(Y)$
ϑ
$X/do(X)$
$do(Y)$
Let $\mathcal{X} = \sum X_t$ or $\int do(X_t)dt$
and $\mathcal{Y} = \oint do(Y_\tau)d\tau$
Causation
ϑ
$\mathcal{X}$
$\mathcal{Y}$

Figure 2: The relation-first symbolic definition of causal relationship versus mere correlation.

A statistical model is typically defined through a function $f(x \mid \theta)$ that represents how a parameter θ is *functionally related* to potential outcomes x of a random variable X Ly et al. (2017). For instance, the coin flip model is also known as the Bernoulli distribution $f(x \mid \theta) = \theta^x (1-\theta)^{1-x}$ with $x \in \{0, 1\}$, which relates the coin's propensity (i.e. its inherent possibility) θ to X = "land heads to the potential outcomes". Formally, given a known θ, the *functional relationship* f yields a probability density function (pdf) as $p_\theta(x) = f(x \mid \theta)$, according to which, X is distributed and denoted as $X \sim f(x; \theta)$. The Fisher Information $\mathcal{I}_X(\theta)$ of X about

θ is defined as $\mathcal{I}_X(\theta) = \int_{\{0,1\}} (\frac{d}{d\theta} log(f(x \mid \theta))^2 p_\theta(x) dx$, with the purpose of building models on the observed x data being to obtain this information. For clarity, we refer to this initial perspective of understanding functional models as the ***relation-first*** *principle*.

In practice, we do not limit all functions to pdfs but often shape them for easier understanding. For instance, let $X^n = (X_1, \ldots, X_n)$ represent an n-trial coin flip experiment, while to simplify, instead of considering the random vector X^n, we may only record the number of heads as $Y = \sum_{i=1}^n X_i$. If these n random variables are assumed to be independent and identically distributed (i.i.d.), governed by the identical θ, the distribution of Y (known as binomial) that describes how θ relates to y would be $f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$. In this case, the conditional probability of the raw data, $P(X^n \mid Y = y, \theta) = 1/\binom{n}{y}$ does not depend on θ. This means that once $Y = y$ is given, the conditional probability of X^n becomes independent of θ, although X^n and Y each depend on θ individually. It concludes that no information about θ remains in X^n once $Y = y$ is observed Fisher et al. (1920); Stigler (1973), denoted as $EI(X^n \to Y) = 0$ in the context of relationship modeling. However, in the absence of the i.i.d. assumption and by using a vector $\vartheta = (\theta_1, \ldots, \theta_n)$ to represent the propensity in the n-trial experiment, we find that $EI(X^n \to Y) > 0$ with respect to ϑ. Here, we revisit the foundational concept of Fisher Information, represented as $\mathcal{I}_{X \to Y}(\theta)$, to define:

Definition 1. A relationship denoted as $X \overset{\theta}{\to} Y$ is considered meaningful in the modeling context due to an *informative* ***relation*** θ, where $\mathcal{I}_{X \to Y}(\theta) > 0$, simplifying as $\mathcal{I}(\theta) > 0$.

Specifically, rather than confining within a function $f(;\theta)$ as its parameter, we treat θ as an individual variable to encapsulate the effective information (EI) as outlined by Hoel. Consequently, the *relation-first principle* asserts that a relationship is characterized and identified by a specific θ, regardless of the appearance of its outcome Y, leading to the following inferences:

1. $\mathcal{I}(\theta)$ inherently precedes and is independent of any observations of the outcome, as well as the chosen function f used to describe the outcome distribution $Y \sim f(y;\theta)$.
2. In a relationship represented by $\mathcal{I}(\theta)$, Y is only used to signify its potential outcomes, without any further "information" defined by Y.
3. In AI modeling contexts, a relationship exists as the representation of $\mathcal{I}(\theta)$, which can be stored, applied, and refined across various scenarios to generate observations accordingly.
4. As Y is determined by $\mathcal{I}(\theta)$, X is also governed by the preceding relationships, which can manifest as either observed x data or established representations in the current relationship modeling.

About Relation θ

As emphasized by the Common Cause principle Dawid (1979), "any nontrivial conditional independence between two observables requires a third, mutual cause" Schölkopf et al. (2021). The critical factor here, however, is "nontrivial" rather than "cause" itself. For a system involving X and Y, if their connection (i.e., under what conditions they become independent) is significant enough to call for a description, it must include information beyond the statistical dependencies present in the system. We use an abstract variable θ to signify such an informative connection between X and Y, unnecessarily referring to tangible entities.

Traditionally, descriptions of relationships are constrained by objective notations and focus on "observable states at specific times". For example, quantifying EI requires its attribution to a state-to-state transition probability matrix S Hoel et al. (2013). However, $EI(S)$ cannot be solely defined by S; it must also account for how the current state $s_0 = S$ or action $do(s_0 = S)$ is related to the probability distributions of past and future states, S_P and S_F, respectively. In practice, its distinction from mere statistical dependence relies on manual specification. The rise of representation learning technology offers a shift towards abstracting informative entities beyond mere observational descriptors, irrespective of whether the values representing them possess concrete interpretations. This lays the groundwork for independently conceptualizing $\mathcal{I}(\theta)$.

For a clearer empirical understanding of θ, consider the following example: A sociological study investigates interpersonal ties through consumption data. Bob and Jim, a father-son duo, have consistent expenditures on craft supplies, suggesting the father's influence on the son's hobbies. However, the "father-son" relational information, represented by $\mathcal{I}(\theta)$, exists only in our perception and cannot be deduced directly from the data alone. Traditional object-first approaches require manual labeling of data points to identify specific outcomes. Conversely, in a relation-first modeling paradigm, abstract knowledge is stored and reused in the form of θ representation, facilitating autonomous identifications.

More importantly, the existence of $\mathcal{I}(\theta)$, as well as its potential extraction, are not limited between solely *observational* distributions denoted by variables X and Y, but can extend to the *timing* dimension - For causal relationships, the significant timing-dimensional distribution of the outcome is essential in differentiating them from mere correlations. Specifically, we use $\mathcal{X}$ and $\mathcal{Y}$ to indicate their integration with the timing dimension, and represent a relationship in the general form $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$. Below, we introduce $\mathcal{X}$ for a general discussion on such variables, followed by incorporating $\mathcal{Y}$ to identify the outcome of the relationship model.

About Dynamic Variable $\mathcal{X}$

Definition 2. For a variable $X \in \mathbb{R}^n$ observed as a time sequence $x^t = (x_1, \ldots, x_t)$, a ***dynamic*** variable $\mathcal{X} = \langle X, t \rangle \in \mathbb{R}^{n+1}$ is formulated by integrating t as the *timing dimension.*

Time series data analysis is often referred to as being "spatial-temporal" Andrienko et al. (2003). However, in modeling contexts, "spatial" is interpreted broadly and not limited to physical spatial measurements (e.g., geographic coordinates); thus, we prefer the term "observational". Furthermore, to avoid the implication of "short duration" often associated with "temporal," we use "timing" to represent the dimension t. Unlike the conventional representation in time series $X^t = (X_1, \ldots, X_t)$ with timestamp t, we consider $\mathcal{X}$ holistically as a *dynamic* variable, similarly for $\mathcal{Y} = \langle Y, \tau \rangle \in \mathbb{R}^{m+1}$. The probability distributions of $\mathcal{X}$, as well as $\mathcal{Y}$, span both *observational* and *timing* dimensions simultaneously.

Specifically, $\mathcal{X}$ can be viewed as the integral of X_t or $do(X_t)$ over the timing dimension t within a required range. The necessity to represent it by $do(X_t)$ rather than X_t is referred to as the ***dynamical significance*** of $\mathcal{X}$. Put simply, if $\mathcal{X}$ can be expressed as $\mathcal{X} = \sum_1^t X_t$, then it is equal to $X^t = (X_1, \ldots, X_t)$ in modeling. Otherwise, $\mathcal{X} = \int_{-\infty}^{\infty} do(X_t)dt$ indicates $\mathcal{X}$ to be *dynamically significant*, characterized by dependencies among the states $\{X_{t-1}, X_t\}$ with an unconstrained $t \in (-\infty, \infty)$. In essence, while including the current state X_t, $do(X_t)$ can be seen as additionally integrating the state-dependence from X_{t-1} to X_t, forming a differential unit within the *timing distribution.* This concept aligns with the "state-dependent" and "state-independent" analysis that Hoel discusses as in causal emergence Hoel et al. (2013).

Theorem 1. Timing becomes a necessary *computational dimension* if and only if the variable in question possesses *dynamical significance*, characterized by a ***nonlinear** distribution* across the timing.

In simpler terms, if a distribution over timing t cannot be adequately represented by a function of the form $x_{t+1} = f(x^t)$, then its nonlinearity is significant to be considered. Here, $[t, t+1]$ indicates any predetermined time lag, set at a constant value. RNN models can effectively extract the nonlinearly significant representation of $\mathcal{X}$ from the sequence X^t and autonomously achieve $\mathcal{X} \xrightarrow{\theta} Y$, by leveraging the relational constraint indicated by $\mathcal{I}(\theta)$, i.e., indexing through θ. Conversely, if "predicting" such an irregular timing distribution is crucial, it implies that it has been identified as the causal effect of some underlying reason.

About Dynamic Outcome $\mathcal{Y}$

Theorem 2. In modeling contexts, identifying a relationship $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$ as *Causality*, distinct from mere *Correlation*, depends on the ***dynamical significance** of the outcome* $\mathcal{Y}$ as required by this relationship.

Figure 2 illustrates the distinction between causality and correlation, where an arrow indicates an informative relation and a dashed line means statistical dependence. If conducting the integral operation for both sides of the do-calculus formation $X/do(X) \to Y$ over timing, we can achieve $\mathcal{X} \to \sum_1^\tau Y_\tau$ with the variable $\mathcal{X}$ allowing to be dynamically significant but the outcome $\sum_1^\tau Y_\tau$ certainly not. Essentially, to guarantee $\mathcal{Y}$

presenting in form of sequence $y^\tau = (y_1, \ldots, y_\tau)$ with constant timestamps $\{1, \ldots, \tau\}$, do-calculus manually conducts a differentiation operation on the relational information $\mathcal{I}(\theta)$ to discretize the outcome timing distribution. This process is to confirm specific τ values at which y_τ can be identified as the effect of a certain $do(x_t)$ or x_t. Accordingly, the static y_τ value will be defined as either the interventional effect $f_V(do(x_t))$ or the observational effect $f_B(x_t)$, with three criteria in place to maintain conditional independence between these two possibilities, as identified by tangible $\Delta\mathcal{I}(\theta)$ elements (i.e., identifiable $do(x_t) \rightarrow y_\tau$ or $x_t \rightarrow y_\tau$):

$$\mathcal{Y} = f(\mathcal{X}) = \sum_t f_V(do(x_t))\cdot f_B(x_t) = \sum_t \left\{ \begin{array}{ll} f_B(x_t) = y_\tau & \text{with } f_V(do(x_t)) = 1 \text{ (Rule 1)} \\ f_V(do(x_t)) = y_\tau & \text{with } f_B(x_t) = 1 \text{ (Rule 2)} \\ 0 = y_\tau & \text{with } f_V(do(x_t)) = 0 \text{ (Rule 3)} \\ \text{otherwise} & \text{not identifiable} \end{array} \right\} = \sum_\tau y_\tau$$

Particularly, the proposed *dynamic* notations $\mathcal{X} = \langle X, t \rangle$ and $\mathcal{Y} = \langle Y, \tau \rangle$ offer advantages in two respects: 1) the concept of $do(Y_\tau)$ can be introduced with a computable τ indicating its possible timing, which cannot stand in the context of causal inference, and 2) by incorporating the timing dimension, $\mathcal{X}$ and $\mathcal{Y}$ transcend the conventional notion of "timestamps", eliminating the need to distinguish between "past and future".

Definition 3. A ***causality*** characterized by a *dynamically significant outcome* $\mathcal{Y}$ can encompass multiple ***causal components***, represented by $\vartheta = (\vartheta_1, \ldots, \vartheta_T)$. Each ϑ_τ with $\tau \in \{1, \ldots, T\}$ identifies a timing dimension τ to accommodate the corresponding ***outcome component*** $\mathcal{Y}_\tau$.
The overall outcome is denoted as $\mathcal{Y} = \sum_{\tau=1}^{T} \mathcal{Y}_\tau = \sum_{\tau=1}^{T} \int do(Y_\tau) d\tau$, simplifying to $\oint do(Y_\tau) d\tau$.

Definition 3 is grounded in the relation-first principle, emphasizing causality as indicated by ϑ (different from θ), which determines the appearance of outcome $\mathcal{Y}$ to be dynamically significant. Specifically, within a general relationship $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$, the dynamic outcome $\mathcal{Y}$ showcases its capability as a variable to encompass nonlinear distribution over timing, whereas $\mathcal{X} \xrightarrow{\vartheta} \mathcal{Y}$ confirms such nature of this relationship.

According to Theorem 1, adding timing as a computational dimension is only meaningful for relationships as causality. Simplified, if a relationship model is expressed as $f(\mathcal{X}; \theta) = Y^\tau = (Y_1, \ldots, Y_\tau)$, it is equal to a model $f(\mathcal{X}; \theta) = Y$ with a static outcome, but applied τ times in sequence, implying that $\mathcal{X} \xrightarrow{\theta} Y$ can adequately represent the modeled relationship. Frequently, it goes unnoticed that incorporating a sequence $X^t = (X_1, \ldots, X_t)$ in modeling does not mean the t-dimension is computationally active, where t serves as a fixed constant, lacking computational flexibility. The same way also applies to Y^τ.

However, once including the "possible timing" τ with computable values, it becomes necessary to account for the potential multiple components of $\mathcal{Y}$ that dynamically develop separately over their own timing. For a simpler understanding, let's revisit the example of "storm causes flooding." Suppose $\mathcal{X}$ represents the storm, and for each watershed, ϑ encapsulates the effects of $\mathcal{X}$ determined by its unique hydrological conditions. Let $\mathcal{Y}_2$ denote the water levels observed over an extended period, such as the next 30 days, if without any flood prevention. Let $\mathcal{Y}_1$ indicate the daily variations in water levels (measured in ±cm to reflect increases or decreases) resulting from flood-prevention efforts. In this case, ϑ can be considered in two components: $\vartheta = (\vartheta_1, \vartheta_2)$, separately identifying $\tau = 1$ and $\tau = 2$.

Specifically, historical records of disasters without flood prevention could contribute to training ϑ_2, based on which, ϑ_1 can be trained using recent records of flood prevention. Even if their hydrological conditions are not exactly the same, AI can represent such relational difference between ϑ_1 and ϑ_2. This is because the capability of computing over timing dimensions empowers AI to extract common relational information from different dynamics. From AI's standpoint, regardless of whether the flood crest naturally occurs on the 7th day or is dispersed over the subsequent 30 days, both $\mathcal{Y}_2$ and $(\mathcal{Y}_1 + \mathcal{Y}_2)$ are linked to $\mathcal{X}$ by the same volume of water introduced by $\mathcal{X}$. In other words, while AI deals with the computations over timing, discerning what qualifies as a "disaster" remains humans' determination.

Conversely, in traditional modeling, ϑ is often viewed as another causal object, named a "confounder" (i.e., common cause) of $\mathcal{X}$ and $\mathcal{Y}$. Therefore, when ϑ_1 and ϑ_2 represent varied conditions, their difference is considered as an unknown object, and naturally requires manual adjustments on the data or even the experiments, to achieve identical ϑ_1 and ϑ_2 ensuring comparable outcome sequences Y_1^τ and Y_2^τ. For example, in clinical trials, patient groups receiving different treatments must be "randomized" similarly to maintain comparability. This is essentially for eliminating the dynamical significance of the outcomes.

About Dependence ϕ between Causal Components

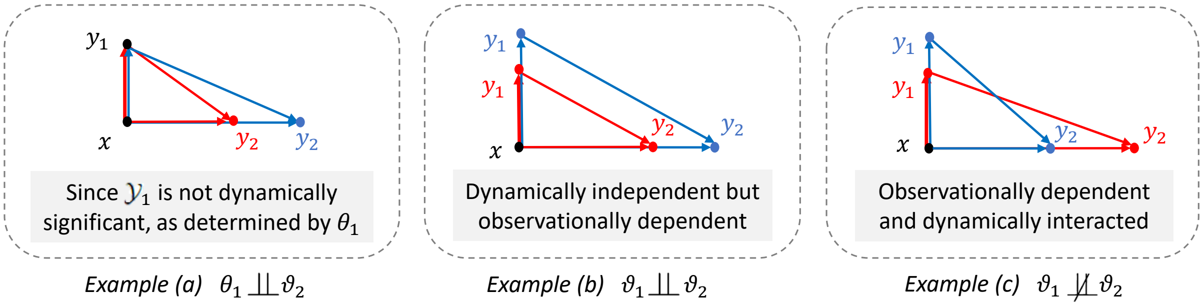
Based on Definition 3, the causal emergence phenomenon in Figure 1(b) can be clearly explained as in (c). The two causal components ϑ_1 and ϑ_2 result in dynamic outcome components $do(Y)_1$ and $do(Y)_2$, acting as differentiated $\mathcal{Y}_1$ and $\mathcal{Y}_2$ respectively. As $EI(\phi) = 0$, no further relational information exists between $do(Y)_1$ and $do(Y)_2$. However, as dynamics, their statistical dependence can be dynamically significant.

Theorem 3. Sequential causal modeling is required, if the ***dependence*** between causal components, represented by ϕ, has dynamically significant impact on the outcome components.

The sequential modeling procedure is used in the analysis of the "flooding" example, where the training of ϑ_1 relies on the previously established ϑ_2 to form meaningful representations. That is, the dependence between ϑ_1 and ϑ_2 represented by ϕ has impacts on the timing dimension of their dynamic outcomes $\mathcal{Y}_1$ and $\mathcal{Y}_2$. As a result, the timing-dimensional distributions of $\mathcal{Y}_1$ and $\mathcal{Y}_2$ must be determined sequentially, depending on the meaningful $\mathcal{I}(\vartheta_1 \mid \vartheta_2)$ or $\mathcal{I}(\vartheta_2 \mid \vartheta_1)$ tailored to specific applications.

In Figure 1(c), suppose $\mathcal{Y}_2$ depends on $\mathcal{Y}_1$, the modeling process could be in two-step: $\mathcal{Y}_1 = f_1(\mathcal{X}; \vartheta_1)$ followed by $\mathcal{Y}_2 = f_2(\mathcal{X} \mid \mathcal{Y}_1; \vartheta_2)$. Remarkably, according to the adopted perspective, its informative explanation can be in different ways. Under the relation-first principle (i.e., in the creator's perspective), $\mathcal{I}(\vartheta) = \mathcal{I}(\vartheta_1) + \mathcal{I}(\vartheta_2) = 2\mathcal{I}(\vartheta_1) + \mathcal{I}(\vartheta_2 \mid \vartheta_1)$ encapsulates all information needed to "create" the outcome $\mathcal{Y} = \mathcal{Y}_1 + \mathcal{Y}_2$, with $\mathcal{I}(\phi) = 0$ uninformatively indicating a dependence. While adopting the traditional information expression (i.e., from the observer's perspective), ϑ_1 and ϑ_2 simply denote functional parameters without *observational information* associated, and accordingly, we have $\mathcal{I}(\phi \mid \mathcal{Y}_1) = \mathcal{I}(\mathcal{Y}_2) - \mathcal{I}(\mathcal{Y}_1) > 0$.

For clarity, we denote $\vartheta_1 \perp\!\!\!\perp \vartheta_2$ to signify the timing-dimensional independence between $\mathcal{Y}_1$ and $\mathcal{Y}_2$, termed as ***dynamical independence***, without altering the conventional understanding within the observational space, like $Y_1 \perp\!\!\!\perp Y_2 \in \mathbb{R}^m$. On the contrary, $\vartheta_1 \not\!\perp\!\!\!\perp \vartheta_2$ implies a ***dynamical interaction*** between $\mathcal{Y}_1$ and $\mathcal{Y}_2$, while "whether *dynamically* independent or interacted" only holds for *dynamically significant* $\mathcal{Y}_1$ and $\mathcal{Y}_2$.


y_1
x
y_2
y_2
Since $\mathcal{Y}_1$ is not dynamically significant, as determined by θ_1
y_1
y_1
x
y_2
y_2
Dynamically independent but observationally dependent
y_1
y_1
x
y_2
y_2
Observationally dependent and dynamically interacted

Example (a) $\theta_1 \perp\!\!\!\perp \vartheta_2$ *Example (b)* $\vartheta_1 \perp\!\!\!\perp \vartheta_2$ *Example (c)* $\vartheta_1 \not\!\perp\!\!\!\perp \vartheta_2$

Figure 3: Examples for understanding dynamical dependence, with the observational dependence from $\mathcal{Y}_1$ to $\mathcal{Y}_2$ displayed as the vector $\overrightarrow{y_1 y_2}$. Red and blue indicate two different instances.

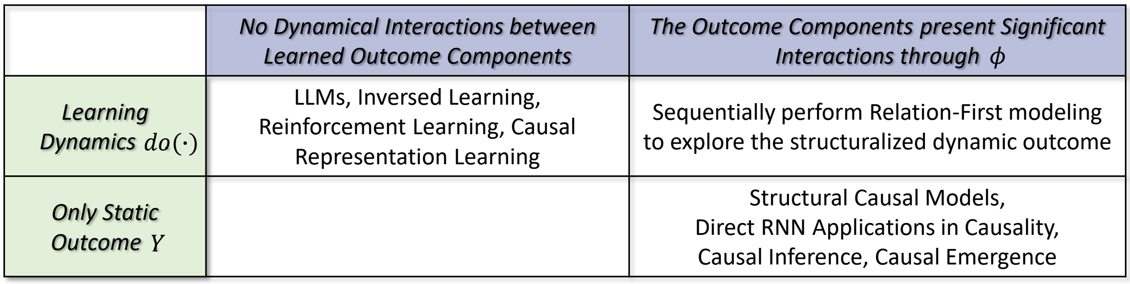
For better illustration, we draw the examples in Figure 3, by upgrading the conventional Directed Acyclic Graph (DAG) in two aspects: 1) nodes represent static values of data instances (i.e., data points or variable realizations), and 2) edge lengths reflect the timespan to achieve these values. This allows for the visualization of dynamical differences among instances, influenced by the same causal effect (or component), as if captured in a snapshot. For instance, Figure 3(c) indicates ϑ_1 and ϑ_2 are inversely associated about timing, meaning that a quicker achievement of y_1 implies a slower one for y_2.

2.3 Potential Development Toward AGI

As demonstrated, choosing between the observer's or the creator's perspective depends on the questions we are targeting, rather than a matter of conflict. In the former, information is gained from observations and represented by observables; while in the latter, relational information preferentially exists as representing the knowledge we aim to construct in modeling, such that once the model is established, we can use it to deduce outcomes as a description of "possible observations in the future" without direct observation.

Causality questions inherently require the creator's perspective, since "informative observations" cannot exist because of nothing. Empirically, it is reflected as the challenge of specifying outcomes in traditional causal modeling, often referred to as "identification difficulty" Zhang (2012). As mentioned by Schölkopf et al. (2021), "we may need a new learning paradigm" to depart from the i.i.d.-based modeling assumption, which essentially asserts the objects we are modeling exactly exist as how we expect them to. We term this conventional paradigm as *object-first* and have introduced the *relation-first* principle accordingly.

The *relation-first* thinking has emerged in the definition of Fisher Information, as well as in do-calculus that differentiates relational information. Moreover, neural networks with the back-propagation strategy achieved its technological realization. Therefore, it's unsurprising that the advent of AI-based representation learning signifies a turning point in causality modeling. As noted by Rumelhart et al. (1986), "a simple procedure in a general purpose is powerful enough." Sometimes, the resolution of problems is not finding an answer, but an approach that makes the problem disappear (from Ludwig Wittgenstein's philosophical works).



	No Dynamical Interactions between Learned Outcome Components	*The Outcome Components present Significant Interactions through ϕ*
Learning Dynamics $do(\cdot)$	LLMs, Inversed Learning, Reinforcement Learning, Causal Representation Learning	Sequentially perform Relation-First modeling to explore the structuralized dynamic outcome
Only Static Outcome Y		Structural Causal Models, Direct RNN Applications in Causality, Causal Inference, Causal Emergence

Figure 4: The $do(Y)$-Paradox in traditional Causality Modeling vs. modern Representation Learning.

From an engineering standpoint, answering the "what ... if?" (i.e., counterfactual) questions indicates the capacity of predicting $do(Y)$ as structuralized dynamic outcomes. Intriguingly, learning dynamics (i.e., the realization of $do(\cdot)$) and predicting outcomes (i.e., facilitating the role of Y) present a paradox under the traditional learning paradigm, as in Figure 4.

About AI-based Dynamical Learning

Learning from dynamics, in particular, is a significant instinctive ability of humans. Representation learning fulfills optimizing over timing dimension, notably achieving such capabilities, especially in large language models (LLMs) Wes (2023), which evoke discussions about achievements toward AGI Schaeffer et al. (2023). Especially, using meta-learning in LLMs Lake & Baroni (2023) has facilitated clusters of significant dynamic components, showcasing the potential for human-like knowledge generalization capabilities. However, it has also been highlighted that LLMs lack a true understanding of causality Pavlick (2023).

The complexity of causal questions lies in the dynamical interactions within a "possible world", not just in computing individual possibilities, whether they are dynamically significant or not. Instead of a single question, "what ... if?" stands for a self-extending logic, where the "if" condition can be repeatedly applied to modeled possibilities. Hence, causality modeling aims to uncover the unobservable knowledge implied by observable $X/do(X) \rightarrow Y/do(Y)$ phenomenons, to enable its outcome $Y/do(Y)$ as beyond direct observations.

Advanced technologies to autonomously identify outcomes, such as inversed reinforcement learning Arora (2021) and causal representation learning, have blurred the boundary between the roles of variable $X/do(X)$ and outcome $Y/do(Y)$, which are manually maintained in traditional causal inference. Their major focus often lies on the improved efficacy in learning dynamics, yet it is frequently overlooked that the foundational RNN architecture aims to accomplish $do(X) \rightarrow Y$ without addressing $do(Y)$.

Essentially, the significant dynamics (including events, actions, and even deductions) that are autonomously extracted by AI via one-time modelings are dedicated to dynamically meaningful $do(X)$, with Y specifying its meaning. With no interactions, they simply exist as a cluster of individual relationships, with each directed from an observed dynamic $do(X)$ to a static possibility Y. In other words, the potential interactions between dynamics dictate them to be $do(Y)$ components, distinguished from a set of individual $do(X)$ ones.

About Static Outcomes in Causal Inference

The causal inference methodology and structural causal models (SCMs) concentrate on causal structures that account for dynamical interactions. However, the *object-first* paradigm has determined their outcomes to be "objectively existing observations", represented by $Y_\tau \in \mathbb{R}^m$ with predetermined timestamps τ. This inherently requires all potential interactions to follow the single "observed timing", aggregated to be within a "one-time dynamic", leading to "structuralized observables" instead of "structuralized dynamics". As in Figure 1, the overlooked observational information $\mathcal{I}(do(Y)) - \mathcal{I}(Y)$ "emerges" to be an informative relation ϕ in a "possible world", implying indeducible.

Such "causal emergence" requires significant efforts on theoretical and empirical interpretations. Particularly, the unknown relation ϕ is often attributed to the well-known "hidden confounder" problem Greenland et al. (1999); Pearl et al. (2000), linked to the fundamental assumptions of causal sufficiency and faithfulness Sobel (1996). In practice, converting causal knowledge represented by DAGs into operational causal models requires meticulous effort Elwert (2013), where data adjustments and model interpretations often rely on human insight Sanchez et al. (2022); Crown (2019). These accomplishments form the foundation of causal inference's value in the era dominated by statistical analysis, before the advent of neural networks. Currently, leveraging representation learning enables the autonomous extraction of relational information in AI modeling.

About Development of Relation-First Paradigm

As highlighted in Theorem 3, causal learning must be intentionally sequential to achieve structuralized dynamic outcomes. When modeling with prior knowledge, the targeted relation ϑ should be determined first, and then, the sequential input and output data, $x^t = (x_1, \ldots, x_t)$ and $y^\tau = (y_1, \ldots, y_\tau)$, can be chosen to enable AI to purposefully extract $\mathcal{I}(\vartheta)$ between them. While for AI-detected meaningful $do(X)$, we should delve into its significance - questioning "if it suggests a $do(Y)$, what $\mathcal{I}(\vartheta)$ have we extracted?" The gained insights guide the decision on whether and how to proceed with the next round of exploration based on it.

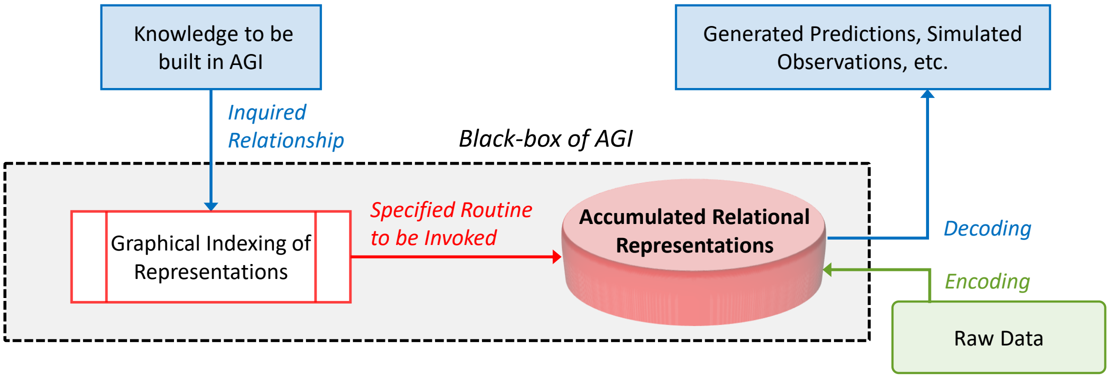

Knowledge to be built in AGI
Generated Predictions, Simulated Observations, etc.
Inquired Relationship
Black-box of AGI
Specified Routine to be Invoked
Graphical Indexing of Representations
Accumulated Relational Representations
Decoding
Encoding
Raw Data

Figure 5: Interfacing with AGI as a Black-box, with human-mediated processes colored in blue. Realizing a usable system demands extensive accumulation and refinement, mirroring the human learning process.

This way, relational representations within the latent space can be accumulated as vital "resources," organized and managed via structured graphical indices, as depicted in Figure 5. This flow mirrors human learning processes Pitt (2022), with these indices corresponding to causal DAGs in our comprehension. Over time, knowledge across different domains can be systematically categorized and accessed like a library system, where the established representations can be continuously optimized and refined across various scenarios.

As humans, we do not need to access latent features directly, encapsulated within the "black box" of the AGI system. The indexing process translates the inquired relationship into a specific input-output routine of representations, guiding the data reconstruction process to generate observations that humans understand. Despite convenience, this can place the "intelligence" of computers under more effective control.

3 Modeling Framework in Creator's Perspective

Under the traditional i.i.d.-based framework, questions must be addressed individually within their respective modeling processes, even when they share similar underlying knowledge. This is due to each modeling process harboring incorrect premises about the objective reality it faces, which often goes unnoticed because of conventional *object-first* thinking. This fundamental issue is further exposed by the advanced modeling flexibility afforded by neural networks. Specifically, it is identified as the *model generalizability* challenge by Schölkopf et al. (2021). They introduced the concept of *causal representation* learning, underscoring the importance of prioritizing causal relational information before specifying observables.

Rather than raising new methods, we aim to emphasize the ***shift of perspective*** that enables modeling across the "possible timing space" beyond the solely observational space. When adopting the creator's perspective, a hyper-dimensional space is embraced to accommodate the abstract variables representing the informative relations. The complete view is in Figure 6, where the notion of ω will be introduced later.

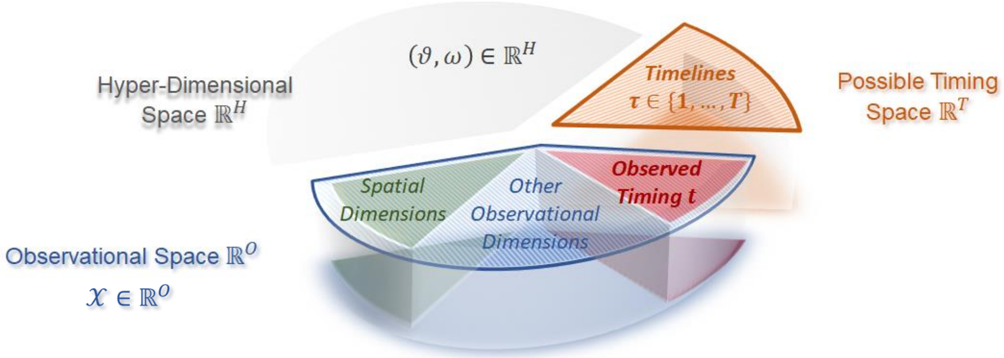

$(\vartheta, \omega) \in \mathbb{R}^H$
Hyper-Dimensional Space $\mathbb{R}^H$
Timelines $\tau \in \{1, \ldots, T\}$
Possible Timing Space $\mathbb{R}^T$
Spatial Dimensions
Other Observational Dimensions
Observed Timing t
Observational Space $\mathbb{R}^O$
$\mathcal{X} \in \mathbb{R}^O$

Figure 6: The modeling framework from the creator's perspective, where $\mathcal{Y} \in \mathbb{R}^{O-1} \cup \mathbb{R}^T$ (with the observed timing t excluded) represents the outcome governed by $\mathcal{I}(\vartheta, \omega)$, without any "observational information" implied. An observer's perspective can be described as $\mathcal{Y} \in \mathbb{R}^{O-1} \cup \tau$ with observational information $\mathcal{I}(\mathcal{Y})$ defined, but without $\mathbb{R}^H$ or $\mathbb{R}^T$ perceived.

Basically, the "creator's perspective" indicates the *distribution* of all possible "observer's perspectives". If considering the latter as a *data point*, then, the former represents the *space* incorporating all possible values of the latter. Thus far, the inexplicable concept of "collapse" in quantum mechanics can be easily explained.

From the creator's perspective, a causal relationship $\mathcal{X} \xrightarrow{\vartheta} \mathcal{Y}$ implies $\mathcal{Y} \in \mathbb{R}^T$, where $\mathbb{R}^T$ is a T-dimensional space with timing $\tau = 1, \ldots, T$ sequentially identifying the T components of $\mathcal{Y}$, whose division depends on the creator's needs. Regardless of how these components are separated, their sum, $\mathcal{Y} = \sum_{\tau=1}^{T} \mathcal{Y}_\tau$, is always governed by $\mathcal{I}(\vartheta)$. However, once the creator "sets up an observer" for this relationship, from the perspective of this "newborn" observer, space $\mathbb{R}^T$ never exists but simply presents as an "observed timeline" τ. In other words, τ has lost its computational flexibility as the "timing dimension" but only exists as a series of constant timestamps. Therefore, the so-called "collapse" indicates such a one-time "perspective shift". Metaphorically speaking, a one-time "collapse" means opening Schrodinger's box for once, while in the modeling context, it implies a one-time modeling computation has been done.

Based on this framework, Theorem 3 can be reinterpreted: Causality modeling is to facilitate "structuralized collapses" within $\mathbb{R}^T$ from the creator's perspective. Notably, for the creator, the concept of $\mathbb{R}^T$ extends beyond representing a single relationship and can encompass "structuralized relationships" by adopting a more macro-level perspective. Accordingly, we introduce the following definitions.

Definition 4. A causal relation ϑ can be defined as ***micro-causal*** if an extraneous relation ω exists with $\mathcal{I}(\omega) \not\subseteq \mathcal{I}(\vartheta)$, such that incorporating ω can form a new, ***macro-causal*** relation, denoted by (ϑ, ω). The process of incorporating ω is referred to as a ***generalization***.

Definition 5. From the creator's perspective, the ***macro-level*** possible timing space $\mathbb{R}^T = \sum_{\tau=1}^{T} \mathbb{R}^\tau$ is constructed by aggregating each ***micro-level*** space $\mathbb{R}^\tau$, where $\tau \in \{1, \ldots, T\}$ indicates the ***timeline*** that houses the sequential timestamps by adopting the observer's perspective for $\mathbb{R}^\tau$.

Notably, both the macro-level possible timing space $\mathbb{R}^T$ and the hyper-dimensional space $\mathbb{R}^H$ are considered "spaces of thought", without a notion of "dimensionality", nor specific "distributions" defined within them. In fact, when we discuss a potential "computation", the observer's perspective has already been set up, from which, the micro-level space $\mathbb{R}^\tau$ (or a cluster of such spaces $\{\mathbb{R}^\tau\}$) has been determined and "ready to be collapsed" by using the approaches in the discussion. In the philosophical context, the concept of timeline τ within a "thought space" $\mathbb{R}^T$ has been introduced as "relative timing" Wulf et al. (1994); Shea et al. (2001), compared to the "absolute timing" t. Furthermore, computations regarding τ in the modeling context can refer to the well-known Granger causality method Granger (1993).

3.1 Hierarchical Levels by ω

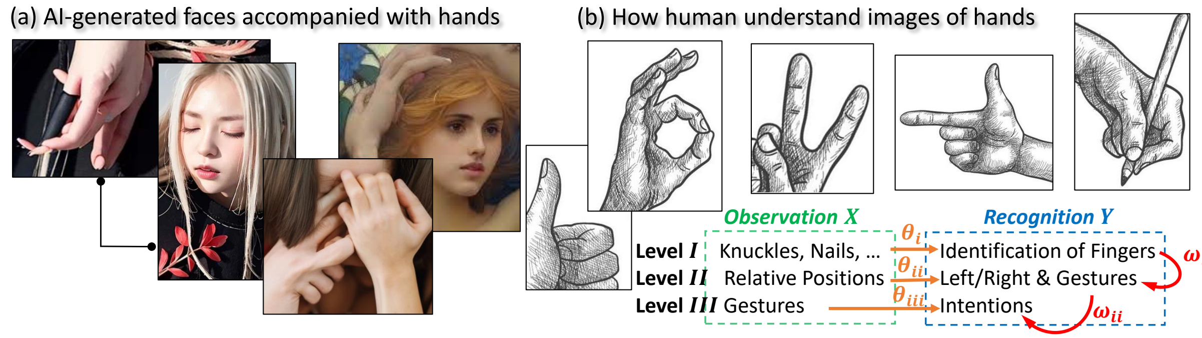
As illustrated in Figure 1, the "causal emergence" phenomenon is due to adopting different perspectives, rather than truly incorporating new relational information. Specifically, we employ the terms "micro-causal" and "macro-causal" to identify such an information gain, which defines a causal *generalization* process as in Definition 4, and the inverse reduction process is called *individualization*. In the modeling context, the ***generalizability*** of an established micro-causal model $f(;\vartheta)$ can be considered as its capability to be reused without reducing the representation of $\mathcal{I}(\vartheta)$, when modeling its macro-causality.

The information gained from $\mathcal{I}(\vartheta)$ to $\mathcal{I}(\vartheta, \omega)$ implies potential incorporations of new observables, which could be new causes, new outcome components, or even both. In general, incorporating ω can introduce a new ***hierarchical level***, which in turn raises the requirements for the generalizability of causal models.

For example, family incomes X influence grocery shopping frequencies Y through the relation θ, denoted as $X \xrightarrow{\theta} Y$, *logically* considered as a causal relationship (without such significance in modeling). The relation θ could vary across countries due to the cultural differences denoted by ω. This raises two hierarchical levels: the global-level relation θ, and the country-level conditional relation $(\theta \mid \omega)$. Although ω does not serve as a modeling objective, it acts as a critical condition, such that the overall information $\mathcal{I}(\theta, \omega) = \mathcal{I}(\theta \mid \omega) + \mathcal{I}(\omega)$ is still required. If adopting the observer's perspective, an additional observable, the country Z, can be incorporated as a new cause, which jointly influences Y with X.

Addressing hierarchies within knowledge is a common issue in relationship modeling. Specifically, timing distribution hierarchies present significant challenges to traditional causal learning methods, leading to the development of a specialized "group-specific learning" Fuller et al. (2007), which primarily depends on manual identifications. However, this approach is no longer viable in modern AI-based applications, necessitating the adoption of the *relation-first* modeling paradigm. Below, we present two examples to demonstrate this necessity: one is solely observational, and the other involves a causality with timing hierarchy.

Observational Hierarchy Example


(a) AI-generated faces accompanied with hands

(b) How human understand images of hands

Figure 7: AI can generate reasonable faces but treat hands as arbitrary mixtures of fingers, while humans understand observations hierarchically to avoid mess, sequentially indexing through $\{\theta_i, \theta_{ii}, \theta_{iii}\}$.

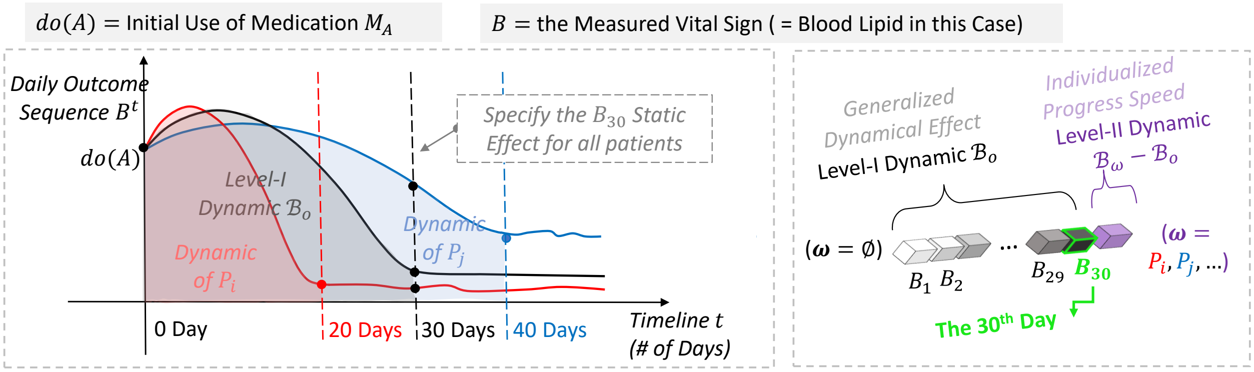
The AI-created personas on social media can have realistic faces but seldom showcase hands, since AI struggles with the intricate structure of hands, instead treating them as arbitrary assortments of finger-like items. Figure 7(a) shows AI-created hands with faithful color but unrealistic shapes, while humans can effortlessly discern hand gestures from the grayscale sketches in (b).

Human cognition intuitively employs informative relations as the *indices* to visit mental representations Pitt (2022). As in (b), this process operates hierarchically, where each higher-level understanding builds upon conclusions drawn at preceding levels. Specifically, Level **I** identifies individual fingers; Level **II** distinguishes gestures based on the positions of the identified fingers, incorporating additional information from our understanding of how fingers are arranged to constitute a hand, denoted by ω_i; and Level **III** grasps the meanings of these gestures from memory, given additional information ω_{ii} from knowledge.

Conversely, AI models often do not distinguish the levels of relational information, instead modeling overall as in a relationship $X \xrightarrow{\theta} Y$ with $\theta = (\theta_i, \theta_{ii}, \theta_{iii})$, resulting a lack of informative insights into ω. However, the hidden information $\mathcal{I}(\omega)$ may not always be essential. For example, AI can generate convincing faces because the appearance of eyes θ_i strongly indicates the facial angles θ_{ii}, i.e., $\mathcal{I}(\theta_{ii}) = \mathcal{I}(\theta_i)$ indicating $\mathcal{I}(\omega_i) = 0$, removing the need to distinguish eyes from faces.

On the other hand, given that X has been fully observed, AI can inversely deduce the relational information using methods such as reinforcement learning Sutton (2018); Arora (2021). In this particular case, when AI receives approval for generating hands with five fingers, it may autonomously begin to derive $\mathcal{I}(\theta_i)$. However, when such hierarchies occur on the timing dimension of a dynamically significant $\mathcal{Y}$, they can hardly be autonomously identified, regardless of whether AI techniques are leveraged.

Timing Hierarchy in Causality Example


$do(A)$ = Initial Use of Medication M_A
B = the Measured Vital Sign (= Blood Lipid in this Case)
Daily Outcome Sequence B^t
$do(A)$
Level-I Dynamic $\mathcal{B}_o$
Dynamic of P_i
Dynamic of P_j
Specify the B_{30} Static Effect for all patients
0 Day
20 Days
30 Days
40 Days
Timeline t (# of Days)
Generalized Dynamical Effect
Level-I Dynamic $\mathcal{B}_o$
Individualized Progress Speed
Level-II Dynamic $\mathcal{B}_\omega - \mathcal{B}_o$
$(\omega = \emptyset)$
B_1 B_2 … B_{29} B_{30}
$(\omega = P_i, P_j, \ldots)$
The 30th Day

(a) Timing Distribution of the Dynamic Outcome $\mathcal{B}_\omega$

(b) Representation of Two-Level Dynamic Outcome

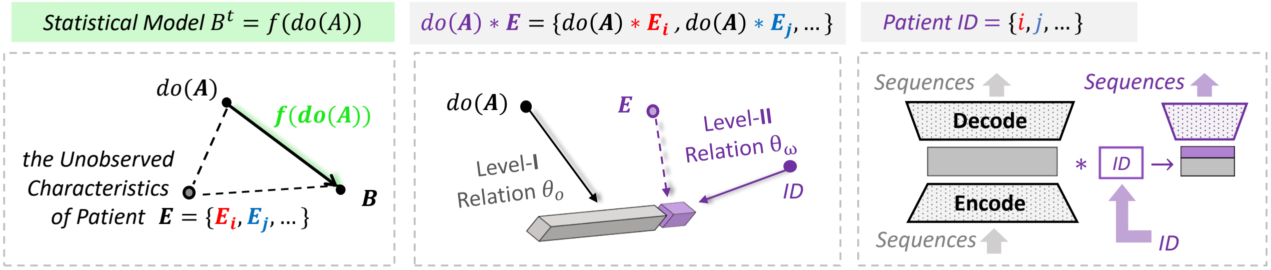
Figure 8: $do(A)$ = the initial use of medication M_A for reducing blood lipid B. By the rule of thumb, the effect of M_A needs around 30 days to fully release ($t = 30$ at the black curve elbow). Patient P_i and P_j achieve the same static effect by 20 and 40 days instead.

In Figure 8, $\mathcal{B}_\omega$ represents the observational sequence $B^t = (B_1, \ldots, B_{30})$ from a group of patients identified by ω. Clinical studies typically aim to estimate the average effect (generalized-level I) on a predetermined day, like $B_{t+30} = f(do(A_t))$. However, our inquiry is indeed the complete level I dynamic $\mathcal{B}_o = \int_{t=1}^{30} do(B_t)dt$, which describes the trend of effect changing over time, without anchored timestamps. To eliminate the level II dynamic from data, a “hidden confounder” is usually introduced to represent their unobserved personal characteristics. Let us denote it by E, and assume E linearly impact $\mathcal{B}_o$, making the level II dynamic $\mathcal{B}_\omega - \mathcal{B}_o$ simply signifying their individualized progress speeds for the same effect $\mathcal{B}_o$.

To accurately represent $\mathcal{B}_o$ with a sequential outcome, traditional methods necessitate an intentional selection or adjustment of training data. This is to ensure the “influence of E” is eliminated from the data, even unavoidable when adopting RNN models. In RNNs, the dynamically significant representation is facilitated only on $do(A)$, while the sequential outcome B^t still requires predetermined timestamps. However, once t

is specified for all patients without the data selection - for example, let $t = 30$ to snapshot B_{30} - bias is inherently introduced, since B_{30} represents the different magnitude of effect $\mathcal{B}_o$ for various patients.

Such hierarchical dynamic outcomes are prevalent in many fields, such as epidemic progression, economic fluctuations, and strategic decision-making. Causal inference typically requires intentional data preprocessing to mitigate inherent biases, including approaches like PSM Benedetto et al. (2018) and backdoor adjustment Pearl (2009), essentially to identify the targeted levels manually. However, they have become impractical due to the modern data volume, and also pose a risk of significant information loss snowballing in structuralized relationship modeling. On the other hand, the significance of timing hierarchies has prompted the development of neural network-based solutions in fields like anomaly detection Wu et al. (2018) to address specific concerns without the intention of establishing a causal modeling framework.


Statistical Model $B^t = f(do(A))$
$do(A) * E = \{do(A) * E_i, do(A) * E_j, \dots\}$
Patient ID $= \{i, j, \dots\}$
$do(A)$
$f(do(A))$
the Unobserved Characteristics of Patient $E = \{E_i, E_j, \dots\}$
B
$do(A)$
E
Level-I Relation θ_o
Level-II Relation θ_ω
ID
Sequences
Sequences
Decode
Encode
$*$
ID
Sequences
ID

(a) DAG with Hidden Confounder (b) Relation-Indexing Disentanglement (c) Latent Space Architecture of (b)

Figure 9: (a) shows the traditional causal DAG for the scenario depicted in Figure 8, (b) disentangles its dynamic outcome in a hierarchical way by indexing through relations, and (c) briefly illustrates the autoencoder architecture for realizing the generalized and individualized reconstructions, respectively.

The concept of "hidden confounder" is essentially elusive, acting more as an interpretational compensation rather than a constructive effort to enhance the model. For example, Figure 9 (a) shows the conventional causal DAG with hidden E depicted. Although the "personal characteristics" are signified, it is not required to be revealed by collecting additional data. This leads to an illogical implication: "Our model is biased due to some unknown factors we don't intend to know." Indeed, this strategy employs a hidden observable to account for the omitted timing-dimensional nonlinearities in statistical models.

As illustrated in Figure 9(b), the associative causal variable $do(A) * E$ remains unknown, unable to form a modelable relationship. On the other hand, *relation-first* modeling approaches only require an observed identifier to index the targeted level in representation extractions, like the patient ID denoted by ω.

3.2 The Generalizability Challenge across Multiple Timelines in $\mathbb{R}^T$

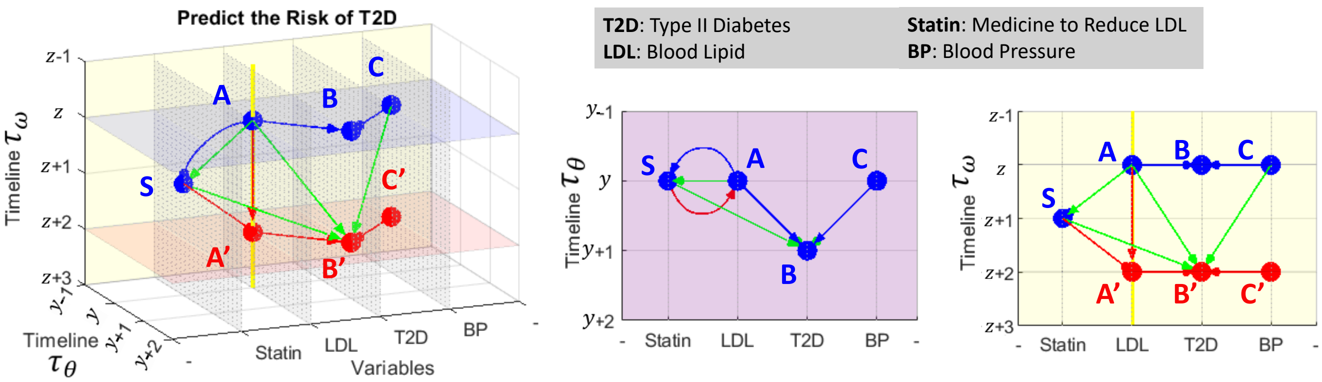


Figure 10: A 3D-view DAG in $\mathbb{R}^{O-1} \cup \mathbb{R}^T$ with two timelines τ_θ and τ_ω. The SCM $B' = f(A, C, S)$ is to evaluate the effect of Statin on reducing T2D risks. On τ_θ, the step $\Delta\tau_\theta$ from y to $(y+1)$ allows A and C to fully influence B; the step $\Delta\tau_\omega$ on τ_ω from $(z+1)$ to $(z+2)$ let S fully release to forward status A to A'.

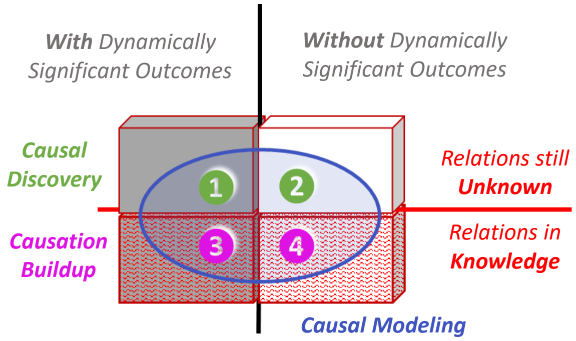
From the creator's perspective, timelines in the macro-level possible timing space $\mathbb{R}^T$ may pertain to different micro-causalities, implying "structuralized" causal relationships. This poses a significant generalizability challenge for traditional structural causal models (SCMs).

The example in Figure 10 showcases a practical scenario in a clinical study. This 3D causal DAG includes two timelines, τ_θ and τ_ω, with the x-axis categorically arranging observables. The upgrades to causal DAGs, as applied in Figure 3, are also adopted here, ensuring that the lengths of the arrows reflect the timespan required to achieve the static values of the observables. Here, the static values are denoted by uppercase letters, representing equal magnitudes of causal effects within the current data population, i.e., the group of patients under analysis. Accordingly, the lengths of the arrows indicate their mean timespans.

We use $\Delta\tau_\theta$ and $\Delta\tau_\omega$ to signify the time steps (i.e., the unit timespans) on τ_θ and τ_ω, respectively. Considering the triangle $SA'B'$, when each unit of effect is delivered from S to A' (taking $\Delta\tau_\omega$), it immediately starts impacting B' through $\overrightarrow{A'B'}$ (with $\Delta\tau_\theta$ required); simultaneously, the next unit of effect begins its generation at S. Under the *relation-first* principle, this dual action requires a two-step modeling process to sequentially extract the dynamic representations on τ_θ and τ_ω. However, in SCM, it is represented by the edge $\overrightarrow{SB'}$ along with a priorly specified timespan from S to B'. This inherently sets the $\Delta\tau_\theta : \Delta\tau_\omega$ ratio based on the current population's performance, freezing the static value represented by B' and fixing the valid shape of the ASB' triangle in this space.

The lack of model generalizability manifests in various ways, depending on the intended scale of generalization. For instance, when focusing on a finer micro-scale causality, the SCM that describes the mean effects for the current population cannot be tailored to individual patients within this population. Conversely, aiming to generalize this SCM to accommodate other populations, or a broader macro-scale causality, may lead to failure because the preset $\Delta\tau_\theta : \Delta\tau_\omega$ ratio lacks universal applicability.

3.3 Fundamental Reliance on Causal Assumptions



	Relational Learning	***Directional Decision***
①	Omitted dynamics are covered by the *Faithfulness Assumption.*	Depending on observational variance, irrelative to causality.
②	Can be aligned with knowledge since no dynamics required.	Maybe suggestive, if observational variance is causally meaningful.
③	Omitted dynamics are covered by *Hidden Confounders* or the *Sufficiency Assumption.*	Predetermined by knowledge.
④	Predetermined by knowledge.	Predetermined by knowledge.

Figure 11: Categories of causal modeling applications. The left rectangular cube indicates all *logically causal* relationships, with the blue circle indicating potentially modelable ones.

Figure 11 categorizes the current causal model applications based on two aspects: 1) if the structure of θ/ϑ is known a priori, they are used for structural causation buildup or causal discovery; 2) depending on whether the required outcome is dynamically significant, they can either accurately represent true causality or not.

Under the conventional modeling paradigm, capturing the significant dynamics within causal outcomes autonomously is challenging. When building causal models based on given prior knowledge, the omitted dynamics become readily apparent. If these dynamics can be specifically attributed to certain unobserved observables, like the node E in Figure 9(a), such information loss is attributed to a hidden confounder. Otherwise, they might be overlooked due to the *causal sufficiency* assumption, which presumes that all potential confounders have been observed within the system. Typical examples of approaches susceptible to these issues are structural equation models (SEMs) and functional causal models (FCMs) Glymour et al. (2019); Elwert (2013). Although state-of-the-art deep learning applications have effectively transformed the discrete structural constraint into continuous optimizations Zheng et al. (2018; 2020); Lachapelle et al. (2019), issues of lack of generalizability still hold Schölkopf et al. (2021); Luo et al. (2020); Ma et al. (2018).

On the other hand, causal discovery primarily operates within the $\mathbb{R}^O$ space and is incapable of detecting dynamically significant causal outcomes. If the interconnection of observables can be accurately specified as the functional parameter θ, there remains a chance to discover informative correlations. Otherwise, mere conditional dependencies among observables are unreliable for causal reasoning, as seen in Bayesian networks Pearl et al. (2000); Peters et al. (2014). Typically, undetected dynamics are overlooked due to the *Causal Faithfulness* assumption, which suggests that the observables can fully represent the underlying causal reality.

Furthermore, the causal directions suggested by the results of causal discovery often lack logical causal implications. Consider X and Y in the optional models $Y = f(X;\theta)$ and $X = g(Y;\phi)$, with predetermined parameters, which indicate opposite directions. Typically, the direction $X \to Y$ would be favored if $\mathcal{L}(\hat{\theta}) > \mathcal{L}(\hat{\phi})$. Let $\mathcal{I}_{X,Y}(\theta)$ denote the information about θ given $\mathbf{P}(X,Y)$. Using $p(\cdot)$ as the density function, the integral $\int_X p(x;\theta)dx$ remains constant in this context. Then:

$$\begin{aligned}\mathcal{I}_{X,Y}(\theta) &= \mathbb{E}[(\frac{\partial}{\partial\theta}\log p(X,Y;\theta))^2 \mid \theta] = \int_Y\int_X (\frac{\partial}{\partial\theta}\log p(x,y;\theta))^2 p(x,y;\theta)dxdy \\ &= \alpha\int_Y (\frac{\partial}{\partial\theta}\log p(y;x,\theta))^2 p(y;x,\theta)dy + \beta = \alpha\mathcal{I}_{Y|X}(\theta)+\beta, \text{with } \alpha,\beta \text{ being constants.}\end{aligned}$$

$$\text{Then, } \hat{\theta} = \arg\max_{\theta} \mathbf{P}(Y \mid X,\theta) = \arg\min_{\theta} \mathcal{I}_{Y|X}(\theta) = \arg\min_{\theta} \mathcal{I}_{X,Y}(\theta), \text{ and } \mathcal{L}(\hat{\theta}) \propto 1/\mathcal{I}_{X,Y}(\hat{\theta}).$$

The inferred directionality indicates how informatively the observational data distribution can reflect the two predetermined parameters. Consequently, such directionality is unnecessarily logically meaningful but could be dominated by the data collection process, with the predominant entity deemed the “cause”, consistent with other existing conclusions Reisach et al. (2021); Kaiser & Sipos (2021).

4 Relation-Indexed Representation Learning (RIRL)

This section introduces a method for realizing the proposed *relation-first* paradigm, referred to as RIRL for brevity. Unlike existing causal representation learning, which is primarily confined to the micro-causal scale, RIRL focuses on facilitating *structural causal dynamics exploration* in the latent space.

Specifically, “relation-indexed” refers to its micro-causal realization approach, guided by the *relation-first* principle, where the indexed representations are capable of capturing the dynamic features of causal outcomes across their timing-dimensional distributions. Furthermore, from a macro-causal perspective, the extracted representations naturally possess high generalizability, ready to be reused and adapted to various practical conditions. This advancement is evident in the structural exploration process within the latent space.

Unlike traditional causal discovery, RIRL exploration spans $\mathbb{R}^O \cup \mathbb{R}^T$ to detect causally significant dynamics without concerns about “hidden confounders”. The causal representations obtained in each round of detection serve as elementary units for reuse, enhancing the flexibility of structural models. This exploration process eventually yields DAG-structured graphical indices, with each input-output pair representing a specific causal routine, readily accessible.

Subsequently, section 4.1 delves into the micro-causal realization to discuss the technical challenges and their resolutions, including the architecture and core layer designs. Section 4.2 introduces the process of “stacking” relation-indexed representations in the latent space, to achieve hierarchical disentanglement at an effect node in DAG. Finally, section 4.3 demonstrates the exploration algorithm from a macro-causal viewpoint.

4.1 Micro-Causal Architecture

For a relationship $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$ given sequential observations $\{x^t\}$ and $\{y^\tau\}$, with $|\overrightarrow{x}| = n$ and $|\overrightarrow{y}| = m$, the relation-indexed representation aims to establish $(\mathcal{X},\theta,\mathcal{Y})$ in the latent space $\mathbb{R}^L$. Firstly, an *initialization* is needed for $\mathcal{X}$ and $\mathcal{Y}$ individually, to construct their latent space representations from observed data sequences. For clarity, we use $\mathcal{H} \in \mathbb{R}^L$ and $\mathcal{V} \in \mathbb{R}^L$ to refer to the latent representations of $\mathcal{X} \in \mathbb{R}^O$ and $\mathcal{Y} \in \mathbb{R}^O$, respectively. The neural network optimization to derive θ is a procedure between $\mathcal{H}$ as input and $\mathcal{V}$ as output. In each iteration, $\mathcal{H}$, θ, and $\mathcal{V}$ are sequentially refined in three steps, until the distance between $\mathcal{H}$ and $\mathcal{V}$ is minimized within $\mathbb{R}^L$, without losing their representations for $\mathcal{X}$ and $\mathcal{Y}$. Consider instances x and y of $\mathcal{X}$ and $\mathcal{Y}$ that are represented by h and v correspondingly in $\mathbb{R}^L$, as in Figure 14. The latent dependency $\mathbf{P}(v|h)$ represents the relational function $f(;\theta)$. The three optimization steps are as follows:

1. Optimizing the cause-encoder by $\mathbf{P}(h|x)$, the relation model by $\mathbf{P}(v|h)$, and the effect-decoder by $\mathbf{P}(y|v)$ to reconstruct the relationship $x \rightarrow y$, represented as $h \rightarrow v$ in $\mathbb{R}^L$.
2. Fine-tuning the effect-encoder $\mathbf{P}(v|y)$ and effect-decoder $\mathbf{P}(y|v)$ to accurately represent y.
3. Fine-tuning the cause-encoder $\mathbf{P}(h|x)$ and cause-decoder $\mathbf{P}(x|h)$ to accurately represent x.

In this process, h and v are iteratively adjusted to reduce their distance in $\mathbb{R}^L$, with θ serving as a bridge to span this distance and guiding the output to fulfill the associated representation $(\mathcal{H}, \theta, \mathcal{V})$. From the perspective of the effect node $\mathcal{Y}$, this tuple represents its component indexing through θ, denoted as $\mathcal{Y}_\theta$.

However, it introduces a technical challenge: for a micro-causality θ, the dimensionality L of the latent space must satisfy $L \geq rank(\mathcal{X}, \theta, \mathcal{Y})$ to provide adequate freedom for computations. To accommodate a structural DAG, this lower boundary can be further enhanced, to be certainly larger than the input vector length $|\overrightarrow{\mathcal{X}}| = t * n$. This necessitates a specialized autoencoder to realize a "higher-dimensional representation", where the accuracy of its reconstruction process becomes significant, and essentially requires *invertibility*.

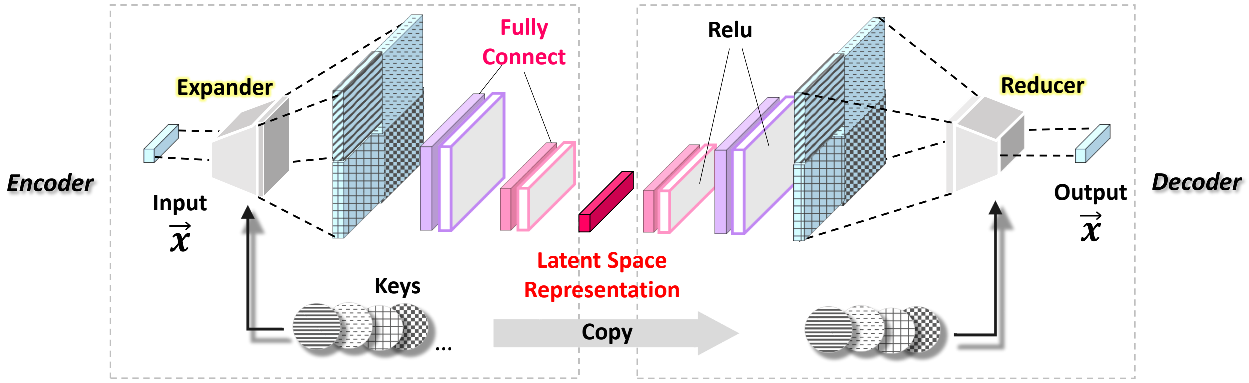


Figure 12: *Invertible* autoencoder architecture for extracting *higher-dimensional* representations.

Figure 12 illustrates the designed autoencoder architecture, featured by a pair of symmetrical layers, named *Expander* and *Reducer* (source code is available [1]). The Expander magnifies the input vector by capturing its higher-order associative features, while the Reducer symmetrically diminishes dimensionality and reverts to its initial formation. For example, the Expander showcased in Figure 12 implements a *double-wise* expansion. Every duo of digits from $\overrightarrow{\mathcal{X}}$ is encoded into a new digit by associating with a random constant, termed the *Key*. This *Key* is generated by the encoder and replicated by the decoder. Such pairwise processing of $\overrightarrow{\mathcal{X}}$ expands its length from $(t * n)$ to be $(t * n - 1)^2$. By concatenating the expanded vectors using multiple *Keys*, $\overrightarrow{\mathcal{X}}$ can be considerably expanded, ready for the subsequent reduction through a regular encoder.

The four blue squares in Figure 12 with unique grid patterns signify the resultant vectors of the four distinct *Keys*, with each square symbolizing a $(t * n - 1)^2$ length vector. Similarly, higher-order expansions, such as *triple-wise* across three digits, can be chosen with adapted *Keys* to achieve more precise reconstructions.

Figure 13 illustrates the encoding and decoding processes within the Expander and Reducer, targeting the digit pair (x_i, x_j) for $i \neq j \in 1, \ldots, n$. The Expander function is defined as $\eta_\kappa(x_i, x_j) = x_j \otimes exp(s(x_i)) + t(x_i)$, which hinges on two elementary functions, $s(\cdot)$ and $t(\cdot)$. The parameter κ represents the adopted *Key* comprising of their weights $\kappa = (w_s, w_t)$. Specifically, the Expander morphs x_j into a new digit y_j utilizing

---

[1] https://github.com/kflijia/bijective_crossing_functions/blob/main/code_bicross_extracter.py

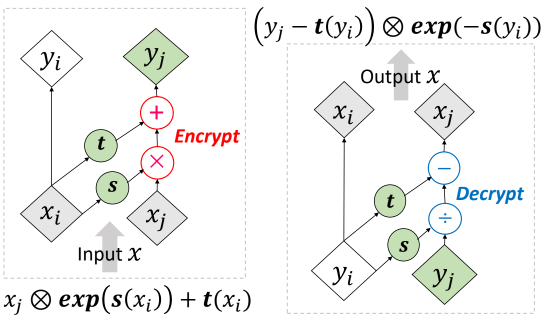

y_i
y_j
+
Encrypt
t
×
s
x_i
x_j
Input x
$x_j \otimes \boldsymbol{exp}(\boldsymbol{s}(x_i)) + \boldsymbol{t}(x_i)$
$(y_j - \boldsymbol{t}(y_i)) \otimes \boldsymbol{exp}(-\boldsymbol{s}(y_i))$
Output x
x_i
x_j
−
Decrypt
t
÷
s
y_i
y_j

Figure 13: Expander (left) and Reducer (right).

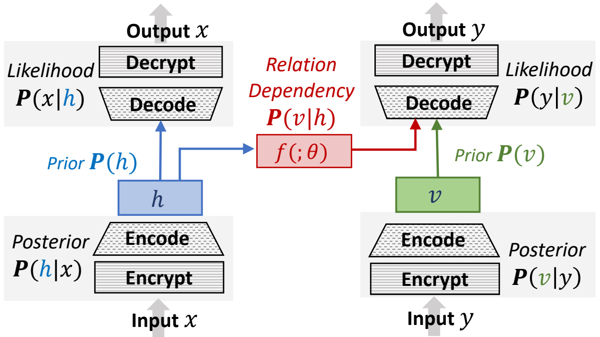

Output x
Likelihood $\boldsymbol{P}(x|h)$
Decrypt
Decode
Relation Dependency $\boldsymbol{P}(v|h)$
$f(;\theta)$
Prior $\boldsymbol{P}(h)$
h
Posterior $\boldsymbol{P}(h|x)$
Encode
Encrypt
Input x
Output y
Decrypt
Decode
Likelihood $\boldsymbol{P}(y|v)$
Prior $\boldsymbol{P}(v)$
v
Encode
Encrypt
Posterior $\boldsymbol{P}(v|y)$
Input y

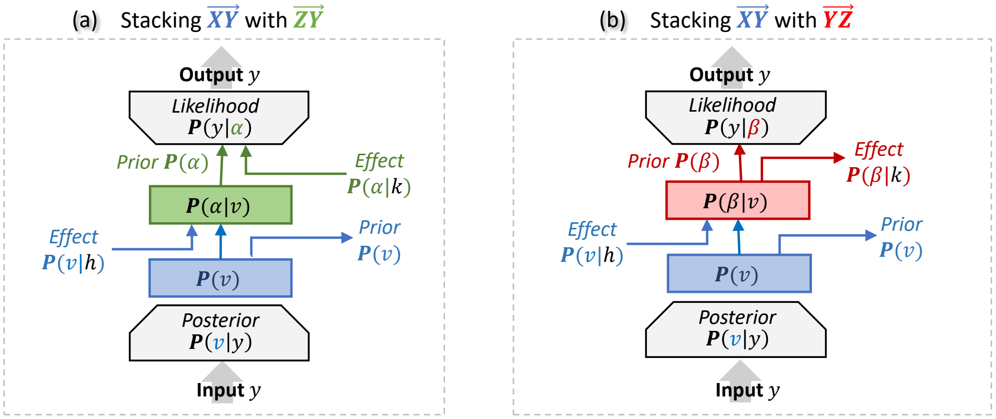
Figure 14: Micro-Causal architecture.

x_i as a chosen attribute. In contrast, the Reducer symmetrically performs the inverse function η_κ^{-1}, defined as $(y_j - t(y_i)) \otimes exp(-s(y_i))$. This approach circumvents the need to compute s^{-1} or t^{-1}, thereby allowing more flexibility for nonlinear transformations through $s(\cdot)$ and $t(\cdot)$. This is inspired by the groundbreaking work in Dinh et al. (2016) on invertible neural network layers employing bijective functions.

4.2 Stacking Relation-Indexed Representations

In each round of detection during the macro-causal exploration, a micro-causal relationship will be selected for establishment. Nonetheless, the cause node in it may have been the effect node in preceding relations, e.g., the component $\mathcal{Y}_\theta$ may already exist at $\mathcal{Y}$ when $\mathcal{Y} \to \mathcal{Z}$ is going to be established. This process of conditional representation buildup is referred to as "stacking".

For a specific node $\mathcal{X}$, the stacking processes, where it serves as the effect, sequentially construct its hierarchical disentanglement according to the DAG. It requires the latent space dimensionality to be larger than $rank(X) + T$, where T represents the in-degree of node $\mathcal{X}$ in this DAG, as well as its number of components as the dynamic effects. From a macro-causal perspective, T can be viewed as the number of necessary edges in a DAG. While to fit it into $\mathbb{R}^L$, a predetermined L must satisfy $L > rank(\mathbf{X}) + T$, where $\mathbf{X}$ represents the data matrix encompassing all observables. In this study, we bypass further discussions on dimensionality boundaries by assuming L is large enough for exploration, and empirically determine L for the experiments.


(a) Stacking $\overrightarrow{XY}$ with $\overrightarrow{ZY}$

(b) Stacking $\overrightarrow{XY}$ with $\overrightarrow{YZ}$

Output y
Likelihood $\boldsymbol{P}(y|\alpha)$
Prior $\boldsymbol{P}(\alpha)$
Effect $\boldsymbol{P}(\alpha|k)$
$\boldsymbol{P}(\alpha|v)$
Effect $\boldsymbol{P}(v|h)$
Prior $\boldsymbol{P}(v)$
$\boldsymbol{P}(v)$
Posterior $\boldsymbol{P}(v|y)$
Input y
Output y
Likelihood $\boldsymbol{P}(y|\beta)$
Prior $\boldsymbol{P}(\beta)$
Effect $\boldsymbol{P}(\beta|k)$
$\boldsymbol{P}(\beta|v)$
Effect $\boldsymbol{P}(v|h)$
Prior $\boldsymbol{P}(v)$
$\boldsymbol{P}(v)$
Posterior $\boldsymbol{P}(v|y)$
Input y

Figure 15: Stacking relation-indexed representations to achieve hierarchical disentanglement.

Figure 15 illustrates the stacking architectures under two different scenarios within a three-node system $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. In this figure, the established relationship $\mathcal{X} \to \mathcal{Y}$ is represented by the blue data streams and layers. The scenarios differ in the causal directions between $\mathcal{Y}$ and $\mathcal{Z}$: the left side represents $\mathcal{X} \to \mathcal{Y} \leftarrow \mathcal{Z}$, while the right side depicts $\mathcal{X} \to \mathcal{Y} \to \mathcal{Z}$.

The hierarchically stacked representations allow for flexible input-output combinations to represent different causal routines as needed. For simple exemplification, we use $\mapsto$ to denote the input and output layers in the stacking architecture. On the left, $\mathbf{P}(v|h) \mapsto \mathbf{P}(\alpha)$ represents the $\mathcal{X} \rightarrow \mathcal{Y}$ relationship, while $\mathbf{P}(\alpha|k)$ implies $\mathcal{Z} \rightarrow \mathcal{Y}$. Conversely, on the right, $\mathbf{P}(v) \mapsto \mathbf{P}(\beta|k)$ denotes the $\mathcal{Y} \rightarrow \mathcal{Z}$ relationship with $\mathcal{Y}$ as the input. Meanwhile, $\mathbf{P}(v|h) \mapsto \mathbf{P}(\beta|k)$ captures the causal sequence $\mathcal{X} \rightarrow \mathcal{Y} \rightarrow \mathcal{Z}$.

4.3 Exploration Algorithm in the Latent Space

Algorithm 1: RIRL Exploration

```
Result: ordered edges set $\mathbf{E} = \{e_1, \dots, e_n\}$
$\mathbf{E} = \{\}$ ; $N_R = \{n_0 \mid n_0 \in N, Parent(n_0) = \varnothing\}$ ;
while $N_R \subset N$ do
    $\Delta = \{\}$ ;
    for $n \in N$ do
        for $p \in Parent(n)$ do
            if $n \notin N_R$ and $p \in N_R$ then
                $e = (p, n)$;
                $\beta = \{\}$;
                for $r \in N_R$ do
                    if $r \in Parent(n)$ and $r \neq p$ then
                        $\beta = \beta \cup r$
                    end
                end
                $\delta_e = K(\beta \cup p, n) - K(\beta, n)$;
                $\Delta = \Delta \cup \delta_e$;
            end
        end
    end
    $\sigma = argmin_e(\delta_e \mid \delta_e \in \Delta)$;
    $\mathbf{E} = \mathbf{E} \cup \sigma$;  $N_R = N_R \cup n_\sigma$;
end
```

$G = (N, E)$	graph G consists of N and E
N	the set of nodes
E	the set of edges
N_R	the set of reachable nodes
$\mathbf{E}$	the list of discovered edges
$K(\beta, n)$	KLD metric of effect $\beta \rightarrow n$
β	the cause nodes
n	the effect node
δ_e	KLD Gain of candidate edge e
$\Delta = \{\delta_e\}$	the set $\{\delta_e\}$ for e
n,p,r	notations of nodes
e,σ	notations of edges

Algorithm 1 outlines the heuristic exploration procedure among the initialized representations of nodes. We employ the Kullback-Leibler Divergence (KLD) as the optimization criterion to evaluate the similarity between outputs, such as the relational $\mathbf{P}(v|h)$ and the prior $\mathbf{P}(v)$. A lower KLD value indicates a stronger causal strength between the two nodes. Additionally, we adopt the Mean Squared Error (MSE) as another measure of accuracy. Considering its sensitivity to data variances Reisach et al. (2021), we do not choose MSE as the primary criterion.

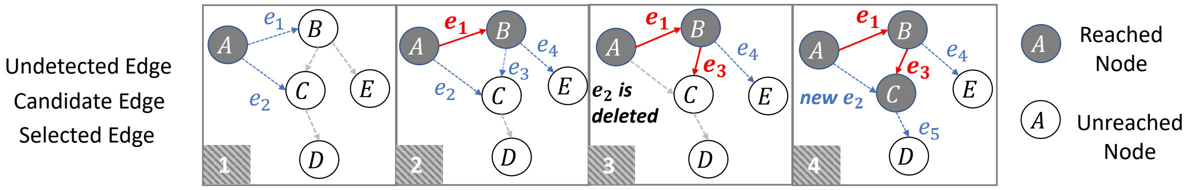


Figure 16: An illustrative example of a detection round in latent space during the RIRL exploration process.

By adopting a macro-causal viewpoint, the example in Figure 16 showcases the process of stacking a new representation for the selected edge, encompassing four primary steps: In Step 1, two edges, e_1 and e_3, have been selected in previous detection rounds. In Step 2, e_1, having been selected, becomes the preceding effect at node B for the next round. In Step 3, with e_3 selected in the new round, the candidate edge e_2 from A to C must be deleted and rebuilt since e_3 alters the conditions at C. Step 4 depicts the resultant structure.

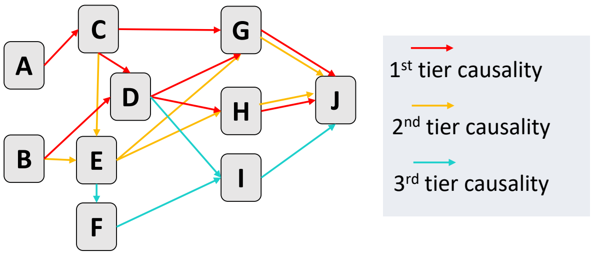
5 RIRL Exploration Experiments

In the experiments, our objective is to evaluate the proposed RIRL method from three perspectives: 1) the performance of the higher-dimensional representation autoencoder, assessed through its reconstruction accuracy; 2) the effectiveness of hierarchical disentanglement for a specific effect node, as determined by the explored causal DAG; 3) the method's ability to accurately identify the underlying DAG structure through exploration. A comprehensive demonstration of the conducted experiments is available online[2]. However, it is important to highlight two primary limitations of the experiments, which are detailed as follows:

Firstly, as an initial realization of the *relation-first* paradigm, RIRL struggles with modeling efficiency, since it requires a substantial amount of data points for each micro-causal relationship, making the heuristic exploration process slow. The dataset used is generated synthetically, thus providing adequate instances. However, current general-use simulation systems typically employ a single timeline to generate time sequences - It means that interactions between dynamic outcome components across multiple timelines cannot be formulated. Ideally, real-world data like clinical records would be preferable for validating the macro-causal model's generalizability. Due to practical constraints, we are unable to access such data for this study and, therefore, designate it as an area for future work. The issues of generalization inherent in such data have been experimentally confirmed in prior work Li et al. (2020), which readers may find informative.

Secondly, the time windows for the cause and effect, denoted by n and m, were fixed at 10 and 1, respectively. This arose from an initial oversight in the experimental design stage, wherein the pivotal role of dynamic outcomes was not fully recognized, and our vision was limited by the RNN pattern. While the model can adeptly capture single-hop micro-causality, it struggles with multi-hop routines like $\mathcal{X} \to \mathcal{Y} \to \mathcal{Z}$, since the dynamics in $\mathcal{Y}$ have been discredited by $m = 1$. However, it does not pose a significant technical challenge to expand the time window in future works.

5.1 Hydrology Dataset



ID	Variable Name	Explanation
A	Environmental set I	Wind Speed, Humidity, Temperature
B	Environmental set II	Temperature, Solar Radiation, Precipitation
C	Evapotranspiration	Evaporation and transpiration
D	Snowpack	The winter frozen water in the ice form
E	Soil Water	Soil moisture in vadose zone
F	Aquifer	Groundwater storage
G	Surface Runoff	Flowing water over the land surface
H	Lateral	Vadose zone flow
I	Baseflow	Groundwater discharge
J	Streamflow	Sensors recorded outputs

Figure 17: Hydrological causal DAG: routine tiers organized by descending causality strength.

The employed dataset is from a widely-used synthetic resource in the field of hydrology, aimed at enhancing streamflow predictions based on observed environmental conditions such as temperature and precipitation. In hydrology, deep learning, particularly RNN models, has gained favor for extracting observational representations and predicting streamflow Goodwell et al. (2020); Kratzert et al. (2018). We focus on a simulation of the Root River Headwater watershed in Southeast Minnesota, covering 60 consecutive virtual years with daily updates. The simulated data is from the Soil and Water Assessment Tool (SWAT), a comprehensive system grounded in physical modules, to generate dynamically significant hydrological time series.

Figure 17 displays the causal DAG employed by SWAT, complete with node descriptions. The hydrological routines are color-coded based on their contribution to output streamflow: Surface runoff (the 1st tier) significantly impacts rapid streamflow peaks, followed by lateral flow (the 2nd tier); baseflow dynamics (the 3rd tier) have a subtler influence. Our exploration process aims to reveal these underlying tiers.

5.2 Higher-Dimensional Reconstruction

This test is based on ten observable nodes, each requiring an individual autoencoder for initialing its higher-dimensional representation. Table 1 lists the characteristics of these observables after being scaled (i.e.,

---

[2] https://github.com/kflijia/bijective_crossing_functions.git

Table 1: Characteristics of observables, and corresponding reconstruction performances.

Variable	Dim	Mean	Std	Min	Max	Non-Zero Rate%	RMSE on Scaled	RMSE on Unscaled	BCE of Mask
A	5	1.8513	1.5496	-3.3557	7.6809	87.54	0.093	0.871	0.095
B	4	0.7687	1.1353	-3.3557	5.9710	64.52	0.076	0.678	1.132
C	2	1.0342	1.0025	0.0	6.2145	94.42	0.037	0.089	0.428
D	3	0.0458	0.2005	0.0	5.2434	11.40	0.015	0.679	0.445
E	2	3.1449	1.0000	0.0285	5.0916	100	0.058	3.343	0.643
F	4	0.3922	0.8962	0.0	8.6122	59.08	0.326	7.178	2.045
G	4	0.7180	1.1064	0.0	8.2551	47.87	0.045	0.81	1.327
H	4	0.7344	1.0193	0.0	7.6350	49.93	0.045	0.009	1.345
I	3	0.1432	0.6137	0.0	8.3880	21.66	0.035	0.009	1.672
J	1	0.0410	0.2000	0.0	7.8903	21.75	0.007	0.098	1.088

Table 2: The brief results from the RIRL exploration.

Edge	A→C	B→D	C→D	C→G	D→G	G→J	D→H	H→J	B→E	E→G	E→H	C→E	E→F	F→I	I→J	D→I
KLD	7.63	8.51	10.14	11.60	27.87	5.29	25.19	15.93	37.07	39.13	39.88	46.58	53.68	45.64	17.41	75.57
Gain	7.63	8.51	1.135	11.60	2.454	5.29	25.19	0.209	37.07	-5.91	-3.29	2.677	53.68	45.64	0.028	3.384

normalized), along with their autoencoders' reconstruction accuracies, assessed in the root mean square error (RMSE), where a lower RMSE indicates higher accuracy for both scaled and unscaled data.

The task is challenged by the limited dimensionalities of the ten observables - maxing out at just 5 and the target node, J, having just one attribute. To mitigate this, we duplicate the input vector to a consistent 12-length and add 12 dummy variables for months, resulting in a 24-dimensional input. A double-wise extension amplifies this to 576 dimensions, from which a 16-dimensional representation is extracted via the autoencoder. Another issue is the presence of meaningful zero-values, such as node D (Snowpack in winter), which contributes numerous zeros in other seasons and is closely linked to node E (Soil Water). We tackle this by adding non-zero indicator variables, called *masks*, evaluated via binary cross-entropy (BCE).

Despite challenges, RMSE values ranging from 0.01 to 0.09 indicate success, except for node F (the Aquifer). Given that aquifer research is still emerging (i.e., the 3rd tier baseflow routine), it is likely that node F in this synthetic dataset may better represent noise than meaningful data.

5.3 Hierarchical Disentanglement

Table 3 provides the performance of stacking relation-indexed representations. For each effect node, the accuracies of its micro-causal relationship reconstructions are listed, including the ones from each single cause node (e.g., $B \to D$ or $C \to D$), and also the one from combined causes (e.g., $BC \to D$). We call them "single-cause" and "full-cause" for clarity. We also list the performances of their initialized variable representations on the left side, to provide a comparative baseline. In micro-causal modeling, the effect node has two outputs with different data stream inputs. One is input from its own encoder (as in optimization step 2), and the other is from the cause-encoder, i.e., indexing through the relation (as in optimization step 1). Their performances are arranged in the middle part, and on the right side of this table, respectively.

The KLD metrics in Table 3 indicate the strength of learned causality, with a lower value signifying stronger. Due to the data including numerous meaningful zeros, we have an additional reconstruction for the binary outcome as "whether zero or not", named "mask" and evaluated in Binary Cross Entropy (BCE).

For example, node J's minimal KLD values suggest a significant effect caused by nodes G (Surface Runoff), H (Lateral), and I (Baseflow). In contrast, the high KLD values imply that predicting variable I using D and F is challenging. For nodes D, E, and J, the "full-cause" are moderate compared to their "single-cause" scores, suggesting a lack of informative associations among the cause nodes. In contrast, for nodes G and H, lower "full-cause" KLD values imply capturing meaningful associative effects through hierarchical stacking. The KLD metric also reveals the most contributive cause node to the effect node. For example, the proximity of the $C \to G$ strength to $CDE \to G$ suggests that C is the primary contributor to this causal relationship.

Figure 18 showcases reconstructed timing distributions for the effect nodes J, G, and I in the same synthetic year to provide a straightforward overview of the hierarchical disentanglement performances. Here, black dots represent the ground truth; the blue line indicates the initialized variable representation and the "full-

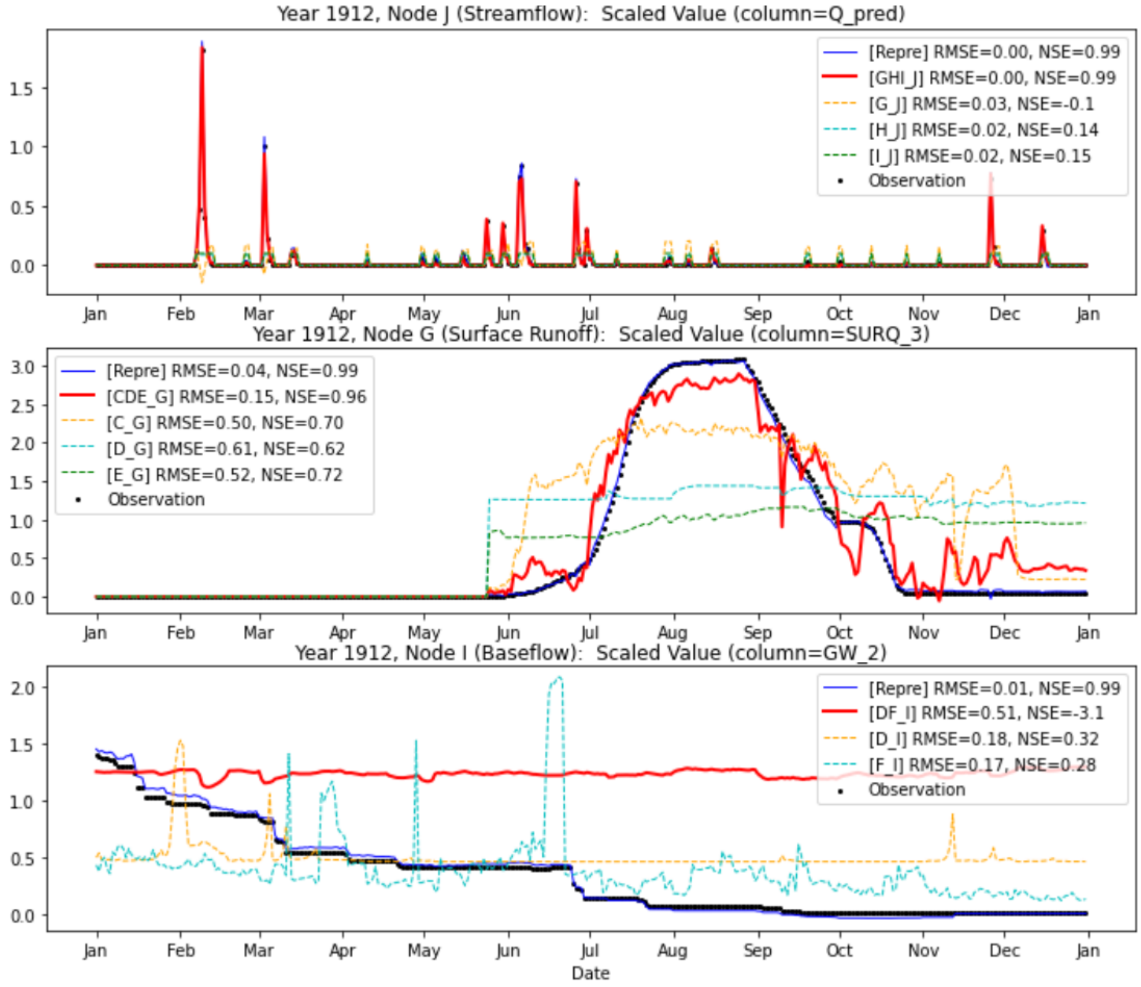


Figure 18: Reconstructed dynamics, via hierarchically stacked relation-indexed representations.

cause" representation generates the red line. In addition to RMSE, we also employ the Nash–Sutcliffe model efficiency coefficient (NSE) as an accuracy metric, commonly used in hydrological predictions. The NSE ranges from $-\infty$ to 1, with values closer to 1 indicating higher accuracy.

The initialized variable representation closely aligns with the ground truth, as shown in Figure 18, attesting to the efficacy of our proposed autoencoder architecture. As expected, the "full-cause" performs better than the "single-cause" for each effect node. Node J exhibits the best prediction, whereas node I presents a challenge. For node G, causality from C proves to be significantly stronger than the other two, D and E.

5.4 DAG Structure Exploration

The first round of detection starts from the source nodes A and B and proceeds to identify their potential edges, until culminating in the target node J. Candidate edges are selected based on their contributions to the overall KLD sum (less gain is better). Table 6 shows the detected order of the edges in Figure 17, accompanied by corresponding KLD sums in each round, and also the KLD gains after each edge is included. Color-coding in the cells corresponds to Figure 17, indicating tiers of causal routines. The arrangement underscores the effectiveness of this latent space exploration approach.

Table 4 in Appendix A displays the complete exploration results, with candidate edge evaluations in each round of detection. Meanwhile, to provide a clearer context about the dataset qualification with respect to underlying structure identification, we also employ the traditional causal discovery method, Fast Greedy Search (FGES), with a 10-fold cross-validation to perform the same procedure as RIRL exploration. The results in Table 5 are available in Appendix A, exhibiting the difficulties of using conventional methods.

Table 3: Performances of micro-causal relationship reconstructions using RIRL, categorized by effect nodes.

Efect Node	Variable Representation (Initialized)			Cause Node	Variable Representation (in Micro-Causal Models)			Relation-Indexed Representation			
	RMSE		BCE		RMSE		BCE	RMSE		BCE	KLD
	on Scaled Values	on Unscaled Values	Mask		on Scaled Values	on Unscaled Values	Mask	on Scaled Values	on Unscaled Values	Mask	(in latent space)
C	0.037	0.089	0.428	A	0.0295	0.0616	0.4278	0.1747	0.3334	0.4278	7.6353
D	0.015	0.679	0.445	BC	0.0350	1.0179	0.1355	0.0509	1.7059	0.1285	9.6502
				B	0.0341	1.0361	0.1693	0.0516	1.7737	0.1925	8.5147
				C	0.0331	0.9818	0.3404	0.0512	1.7265	0.3667	10.149
E	0.058	3.343	0.643	BC	0.4612	26.605	0.6427	0.7827	45.149	0.6427	39.750
				B	0.6428	37.076	0.6427	0.8209	47.353	0.6427	37.072
				C	0.5212	30.065	1.2854	0.7939	45.791	1.2854	46.587
F	0.326	7.178	2.045	E	0.4334	8.3807	3.0895	0.4509	5.9553	3.0895	53.680
G	0.045	0.81	1.327	CDE	0.0538	0.9598	0.0878	0.1719	3.5736	0.1340	8.1360
				C	0.1057	1.4219	0.1078	0.2996	4.6278	0.1362	11.601
				D	0.1773	3.6083	0.1842	0.4112	8.0841	0.2228	27.879
				E	0.1949	4.7124	0.1482	0.5564	10.852	0.1877	39.133
H	0.045	0.009	1.345	DE	0.0889	0.0099	2.5980	0.3564	0.0096	2.5980	21.905
				D	0.0878	0.0104	0.0911	0.4301	0.0095	0.0911	25.198
				E	0.1162	0.0105	0.1482	0.5168	0.0097	3.8514	39.886
I	0.035	0.009	1.672	DF	0.0600	0.0103	3.4493	0.1158	0.0099	3.4493	49.033
				D	0.1212	0.0108	3.0048	0.2073	0.0108	3.0048	75.577
				F	0.0540	0.0102	3.4493	0.0948	0.0098	3.4493	45.648
J	0.007	0.098	1.088	GHI	0.0052	0.0742	0.2593	0.0090	0.1269	0.2937	5.5300
				G	0.0077	0.1085	0.4009	0.0099	0.1390	0.4375	5.2924
				H	0.0159	0.2239	0.4584	0.0393	0.5520	0.4938	15.930
				I	0.0308	0.4328	0.3818	0.0397	0.5564	0.3954	17.410

6 Conclusions

This paper focuses on the inherent challenges of the traditional i.i.d.-based learning paradigm in addressing causal relationships. Conventionally, we construct statistical models as observers of the world, grounded in epistemology. However, adopting this perspective assumes that our observations accurately reflect the "reality" as we understand it, implying that seemingly objective models may actually be based on subjective assumptions. This fundamental issue has become increasingly evident in causality modeling, especially with the rise of applications in causal representation learning that aim to automate the specification of causal variables traditionally done manually.

Our understanding of causality is fundamentally based on the creator's perspective, as the "what...if" questions are only valid within the possible world we conceive in our consciousness. The advocated "perspective shift" represents a transformation from an *object-first* to a *relation-first* modeling paradigm, a change that transcends mere methodological or technical advancements. Indeed, this shift has been facilitated by the advent of AI, particularly through neural network-based representation learning, which lays the groundwork for implementing *relation-first* modeling in computer engineering.

The limitation of the observer's perspective in traditional causal inference prevents the capture of dynamic causal outcomes, namely, the nonlinear timing distributions across multiple "possible timelines". Accordingly, this oversight has led to compensatory efforts, such as the introduction of hidden confounders and the reliance on the sufficiency assumption. These theories have been instrumental in developing knowledge systems across various fields over the past decades. However, with the rapid advancement of AI techniques, the time has come to move beyond the conventional modeling paradigm toward the potential realization of AGI.

In this paper, we present the *relation-first* framework for causality modeling, based on discussions about its philosophical and mathematical underpinnings. Adopting this new framework allows us to significantly simplify or even bypass complex questions. We also introduce the Relation-Indexed Representation Learning (RIRL) method as an initial application of the *relation-first* paradigm, supported by experiments that validate its efficacy.

References

Natalia Andrienko, Gennady Andrienko, and Peter Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, 2003.

Saurabh Arora, Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

Umberto Benedetto, Stuart J Head, Gianni D Angelini, and Eugene H Blackstone. Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, 53(6): 1112–1117, 2018.

William H Crown. Real-world evidence, causal inference, and machine learning. *Value in Health*, 22(5): 587–592, 2019.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

Laurent Dinh, Jascha Sohl, and Samy Bengio. Density estimation using real nvp. *arXiv:1605.08803*, 2016.

Frederick Eberhardt and Lin Lin Lee. Causal emergence: When distortions in a map obscure the territory. *Philosophies*, 7(2):30, 2022.

Felix Elwert. Graphical causal models. *Handbook of causal analysis for social research*, pp. 245–273, 2013.

Ronald Aylmer Fisher et al. 012: A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. 1920.

Ursula Fuller, Colin G Johnson, Tuukka Ahoniemi, Diana Cukierman, Isidoro Hernán-Losada, Jana Jackova, Essi Lahtinen, Tracy L Lewis, Donna McGee Thompson, Charles Riedesel, et al. Developing a computer science-specific learning taxonomy. *ACm SIGCSE Bulletin*, 39(4):152–170, 2007.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Allison E Goodwell, Peishi Jiang, Benjamin L Ruddell, and Praveen Kumar. Debates—does information theory provide a new paradigm for earth science? causality, interaction, and feedback. *Water Resources Research*, 56(2):e2019WR024940, 2020.

Clive WJ Granger. Modelling non-linear economic relationships. *OUP Catalogue*, 1993.

Sander Greenland, Judea Pearl, and James M Robins. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

Erik P Hoel. When the map is better than the territory. *Entropy*, 19(5):188, 2017.

Erik P Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

Yimin Huang, Marco Valtorta. Pearl's calculus of intervention is complete. *arXiv:1206.6831*, 2012.

Marcus Kaiser and Maksim Sipos. Unsuitability of notears for causal graph discovery. *arXiv:2104.05441*, 2021.

Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. Rainfall–runoff modelling using lstm networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, pp. 1–7, 2023.

Jia Li, Xiaowei Jia, Haoyu Yang, Vipin Kumar, Michael Steinbach, and Gyorgy Simon. Teaching deep learning causal effects improves predictive performance. *arXiv preprint arXiv:2011.05466*, 2020.

Yunan Luo, Jian Peng, and Jianzhu Ma. When causal inference meets deep learning. *Nature Machine Intelligence*, 2(8):426–427, 2020.

Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul PPP Grasman, and Eric-Jan Wagenmakers. A tutorial on fisher information. *Journal of Mathematical Psychology*, 80:40–55, 2017.

Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature methods*, 15(4):290–298, 2018.

Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

Tshilidzi Marwala. *Causality, correlation and artificial intelligence for rational decision making.* World Scientific, 2015.

James Massey et al. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pp. 303–305, 1990.

Allen Newell, Herbert A Simon. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, pp. 1975. 2007.

Mohammed Ombadi, Phu Nguyen, Soroosh Sorooshian, and Kuo-lin Hsu. Evaluation of methods for causal discovery in hydrometeorological systems. *Water Resources Research*, 56(7):e2020WR027251, 2020.

Ellie Pavlick. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041, 2023.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

David Pitt. Mental Representation. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

Alexander G Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! varsortability in additive noise models. *arXiv preprint arXiv:2102.13647*, 2021.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *IEEE*, 109(5):612–634, 2021.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Charles H Shea, Gabriele Wulf, Jin-Hoon Park, and Briana Gaunt. Effects of an auditory model on the learning of relative and absolute timing. *Journal of motor behavior*, 33(2):127–138, 2001.

Michael E Sobel. An introduction to causal inference. *Sociological Methods & Research*, 24(3):353–379, 1996.

Stephen M Stigler. Studies in the history of probability and statistics. xxxii: Laplace, fisher, and the discovery of the concept of sufficiency. *Biometrika*, 60(3):439–445, 1973.

Richard S Sutton, Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

Giulio Tononi and Olaf Sporns. Measuring information integration. *BMC neuroscience*, 4:1–20, 2003.

Matej Vuković, Stefan Thalmann. Causal discovery in manufacturing: A structured literature review. *Journal of Manufacturing and Materials Processing*, 6(1):10, 2022.

Gurnee Wes, Tegmark Max. Language models represent space and time, 2023.

Christopher J Wood, Robert W Spekkens. The lesson of causal discovery algorithms for quantum correlations: Causal explanations of bell-inequality violations require fine-tuning. *New Journal of Physics*, 17(3):033002, 2015.

Jia Wu, Weiru Zeng, and Fei Yan. Hierarchical temporal memory method for time-series-based anomaly detection. *Neurocomputing*, 273:535–546, 2018.

Gabriele Wulf, Timothy D Lee, and Richard A Schmidt. Reducing knowledge of results about relative versus absolute timing: Differential effects on learning. *Journal of motor behavior*, 26(4):362–369, 1994.

Haoyan Xu, Yida Huang, Ziheng Duan, Jie Feng, and Pengyu Song. Multivariate time series forecasting based on causal inference with transfer entropy and graph neural network. *arXiv:2005.01185*, 2020.

Kun Zhang, Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

A Appendix: Complete Experimental Results in DAG Structure Exploration Test

Table 4: The Complete Results of RIRL Exploration in the Latent Space. Each row stands for a round of detection, with '#" identifying the round number, and all candidate edges are listed with their KLD gains as below. 1) Green cells: the newly detected edges. 2) Red cells: the selected edge. 3) Blue cells: the trimmed edges accordingly.

# 1	**A→ C**	A→ D	A→ E	A→ F	**B→ C**	B→ D	B→ E	B→ F								
	7.6354	19.7407	60.1876	119.7730	**8.4753**	8.5147	65.9335	132.7717								
# 2	A→ D	A→ E	A→ F	**B→ D**	B→ E	B→ F	C→ D	C→ E	C→ F	C→ G	C→ H	C→ I				
	19.7407	60.1876	119.7730	**8.5147**	65.9335	132.7717	10.1490	46.5876	111.2978	11.6012	39.2361	95.1564				
# 3	**A→ D**	A→ E	A→ F	B→ E	B→ F	**C→ D**	C→ E	C→ F	C→ G	C→ H	C→ I	D→ E	D→ F	D→ G	D→ H	D→ I
	9.7357	60.1876	119.7730	65.9335	132.7717	**1.1355**	46.5876	111.2978	11.6012	39.2361	95.1564	63.7348	123.3203	27.8798	25.1988	75.5775
# 4	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	**C→ G**	C→ H	C→ I	D→ E	D→ F	D→ G	D→ H	D→ I		
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	**11.6012**	39.2361	95.1564	63.7348	123.3203	27.8798	25.1988	75.5775		
# 5	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	C→ H	C→ I	D→ E	D→ F	**D→ G**	D→ H	D→ I	G→ J		
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	39.2361	95.1564	63.7348	123.3203	**2.4540**	25.1988	75.5775	5.2924		
# 6	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	C→ H	C→ I	D→ E	D→ F	D→ H	D→ I	**G→ J**			
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	39.2361	95.1564	63.7348	123.3203	25.1988	75.5775	**5.2924**			
# 7	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	**C→ H**	C→ I	D→ E	D→ F	**D→ H**	D→ I				
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	**39.2361**	95.1564	63.7348	123.3203	**25.1988**	75.5775				
# 8	A→ E	A→ F	B→ E	B→ F	C→ E	C→ F	C→ I	D→ E	D→ F	D→ I	**H→ J**					
	60.1876	119.7730	65.9335	132.7717	46.5876	111.2978	95.1564	63.7348	123.3203	75.5775	**0.2092**					
# 9	**A→ E**	A→ F	B→ E	B→ F	**C→ E**	C→ F	C→ I	D→ E	D→ F	D→ I						
	60.1876	119.7730	65.9335	132.7717	**46.5876**	111.2978	95.1564	63.7348	123.3203	75.5775						
# 10	A→ F	**B→ E**	B→ F	C→ F	C→ I	**D→ E**	D→ F	D→ I	E→ F	E→ G	E→ H	E→ I				
	119.7730	**-6.8372**	132.7717	111.2978	95.1564	**17.0407**	123.3203	75.5775	53.6806	-5.9191	-3.2931	110.2558				
# 11	A→ F	B→ F	C→ F	C→ I	D→ F	D→ I	E→ F	**E→ G**	E→ H	E→ I						
	119.7730	132.7717	111.2978	95.1564	123.3203	75.5775	53.6806	**-5.9191**	-3.2931	110.2558						
# 12	A→ F	B→ F	C→ F	C→ I	D→ F	D→ I	E→ F	**E→ H**	E→ I							
	119.7730	132.7717	111.2978	95.1564	123.3203	75.5775	53.6806	**-3.2931**	110.2558							
# 13	**A→ F**	**B→ F**	**C→ F**	C→ I	**D→ F**	D→ I	**E→ F**	E→ I								
	119.7730	**132.7717**	**111.2978**	95.1564	**123.3203**	75.5775	**53.6806**	110.2558								
# 14	C→ I	D→ I	**E→ I**	**F→ I**												
	95.1564	75.5775	**110.2558**	**45.6490**												
# 15	C→ I	D→ I	**I→ J**													
	15.0222	3.3845	**0.0284**													
# 16	**C→ I**	**D→ I**														
	15.0222	**3.3845**														

Table 5: Average performance of 10-Fold FGES (Fast Greedy Equivalence Search) causal discovery, with the prior knowledge that each node can only cause the other nodes with the same or greater depth with it. An edge means connecting two attributes from two different nodes, respectively. Thus, the number of possible edges between two nodes is the multiplication of the numbers of their attributes, i.e., the lengths of their data vectors.
(All experiments are performed with 6 different Independent-Test kernels, including chi-square-test, d-sep-test, prob-test, disc-bic-test, fisher-z-test, mvplr-test. But their results turn out to be identical.)

Cause Node	A	B		C			D			E			F	G		H	I
True Causation	A→ C	B→ D	B→ E	C→ D	C→ E	C→ G	D→ G	D→ H	D→ I	E→ F	E→ G	E→ H	F→ I	G→ J		H→ J	I→ J
Number of Edges	16	24	16	6	4	8	12	12	9	8	8	8	12	4		4	3
Probability of Missing	0.038889	0.125	0.125	0.062	0.06875	0.039286	0.069048	0.2	0.142857	0.3	0.003571	0.2	0.142857	0.0		0.072727	0.030303
Wrong Causation Times of Wrongly Discovered				C→ F				D→ E	D→ F				F→ G	G→ H	G→ I	H→ I	
				5.6				1.2	0.8				5.0	8.2	3.0	2.8	

Table 6: Brief Results of the Heuristic Causal Discovery in latent space, identical with Table 3 in the paper body, for better comparison to the traditional FGES methods results on this page.
The edges are arranged in detected order (from left to right) and their measured causal strengths in each step are shown below correspondingly.
Causal strength is measured by KLD values (less is stronger). Each round of detection is pursuing the least KLD gain globally. All evaluations are in 4-Fold validation average values. Different colors represent the ground truth causality strength tiers (referred to the Figure 10 in the paper body).

Causation	A→ C	B→ D	C→ D	C→ G	D→ G	G→ J	D→ H	H→ J	C→ E	B→ E	E→ G	E→ H	E→ F	F→ I	I→ J	D→ I
KLD	7.63	8.51	10.14	11.60	27.87	5.29	25.19	15.93	46.58	65.93	39.13	39.88	53.68	45.64	17.41	75.57
Gain	7.63	8.51	1.135	11.60	2.454	5.29	25.19	0.209	46.58	-6.84	-5.91	-3.29	53.68	45.64	0.028	3.384