Relation-Oriented: Toward Causal Knowledge-Aligned AGI

Anonymous authors
Paper under double-blind review

Abstract

The current relationship modeling paradigm, grounded in the observational i.i.d assumption, inherently misaligns with our causal knowledge comprehension due to two vital oversights: 1) the unobservable relations, which lead to undetectable hierarchical levels of knowledge, driving the need for model generalizability; 2) the counterfactual relative timings to support our structural causal reasoning, which lead to inherent biases in models under the current *Observation-Oriented* paradigm. This paper proposes a novel *Relation-Oriented* framework, to reconsider these fundamental questions and unify various confusions surrounding AI-based causal learning, ranging from traditional causal inference to modern language models.

Also, *relation-indexed representation learning* (RIRL) is raised as a baseline implementation method of the proposed new paradigm, alongside comprehensive experiments demonstrating its efficacy in autonomously identifying dynamical effects in relationship modeling.

1 Introduction

The concept of Artificial General Intelligence (AGI) has prompted extensive discussions over the years Newell (2007), with the target toward facilitating human-like causal reasoning and knowledge comprehension in AI systems Marcus (2020). In recent years, the large language models (LLMs) have risen as notable achievements in semantic understanding tasks and accordingly evoked debates about whether LLMs have edged us closer to realizing AGI Rylan (2023). Some studies point to their shortcomings in truly comprehending causality Pavlick (2023), while others argue in favor of LLMs' ability to represent complex spatial and temporal features Wes (2023). Notably, the use of meta-learning in language models has shown potential in achieving human-like generalization capabilities, at least to a certain extent Lake (2023).

These debates are anchored in a fundamental question: What underpins the distinction between two types of generalization? One is how humans generalize learned causal knowledge to diverse scenarios, and another is how AI systems generalize captured associative knowledge among texts and images.

It appears that classical causal inference has offered a clear delineation among causality, correlation, and mere association Pearl et al. (2000); Peters et al. (2017). Moreover, it has provided a robust theoretical groundwork for representing causality in computational models. Based on that, causal learning has been widely utilized and yielded significant contributions to knowledge accumulation in various fields Wood (2015); Vuković (2022); Ombadi (2020). Thus, it is logical to incorporate well-established causal knowledge, often represented as causal DAGs (Directed Acyclic Graphs), into AI model architectures Marwala (2015); Lachapelle et al. (2019). While this integration has greatly enhanced learning efficiency, it has not yet achieved the level of generalizability that constitutes a success Luo (2020); Ma (2018).

This finding likely circles us back to the beginning, as causal inference does not directly bridge the gap between AI models and causal reasoning. However, it does offer a valuable perspective: How would humans conduct causal reasoning based solely on a DAG? While AI is evidently challenged.

Indeed, even within the realm of causal inference, converting DAGs into operational causal models requires rigorous effort Elwert (2013). In various applications, data adjustments and model interpretations are often tailored, relying heavily on human discernment Sanchez (2022); Crown (2019). Challenges include verifying basic causal assumptions Sobel (1996), addressing confounding effects Greenland (1999), and ensuring model interpretability Pearl et al. (2000), among others. These achieved methodologies constitute the cornerstone of the value provided by causal inference. It stands to reason that the answer to our question may be gleaned by examining the challenges causal inference has faced and resolutions adopted.

From an applicational standpoint, Scholkopf (2021) synthesize the development of causal models, emphasizing the crucial role of "causal representations" in achieving AI-based causal models' generalizability across various "levels of knowledge" learning. They propose the potential need for a "new learning paradigm" - an idea we find both logical and thought-provoking. Our current models, ranging from causal to AI, are chiefly based on assumed independent and identically distributed (i.i.d) observations, possibly hindering their ability to achieve generalizable causal learning. Moreover, Zhang (2012) points out the "identification difficulty" when facing nonlinear (i.e., dynamical) effects, an inherent obstacle under the observational i.i.d setting.

For clarity, we designate the prevailing paradigm as ***Observation-Oriented*** modeling. In this study, we propose a novel framework, termed ***Relation-Oriented*** modeling, inspired by the relation-indexing nature of human cognition processes Pitt (2022). Through this new lens, we seek to pinpoint the intrinsic limitations underlying existing modeling approaches. Accordingly, to validate the proposed new paradigm, it must shed light on the array of questions that have emerged from the outset. To encapsulate these queries:

- *Firstly*, causal inference challenges such as confounding effects, dependency on causal assumptions, and interpretative complexities call for a foundational explanation.
- *Secondly*, To integrate causal reasoning within AI models, we need a nuanced understanding of "levels of knowledge," the essential role of causal representation, its relevance to the difficulty of identifying temporally nonlinear effects, and potential resolutions to these issues.
- *Thirdly*, in the context of Large Language Models (LLMs), it is crucial to discern the distinction between the "spatial and temporal" concepts in language understanding versus those in causality comprehension, and critically interpret what meta-learning has accomplished in terms of generalizability.

While these questions may appear disparate, they are intrinsically linked by a fundamental requirement under the observational i.i.d assumption: the prior specification of observables (including their temporal events) in modeling. The specified observational entities serve as the modeling target in solely observational learning tasks (like image recognition). In causal relationship learning, they are priorly identified as causes and effects, with their interrelation acting as the learning objective.

This requirement leads to two ***primary limitations*** in modeling: 1) the inability to account for unobservable relational knowledge, which leads to undetectable hierarchical levels to challenge the model's generalizability, and 2) the prior obligation to identify relational effects along the absolute timing, potentially overlooking the underlying relative timings, which underpin our causal knowledge structure, and leading to inherent biases.

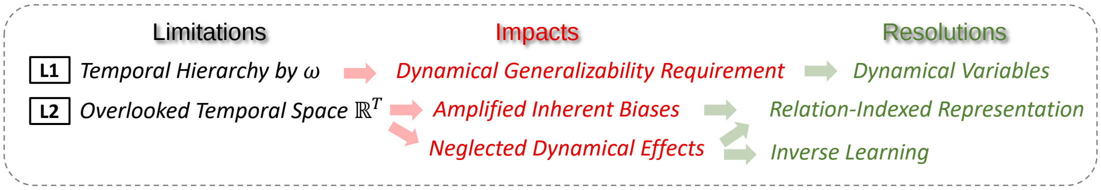

Limitations
Impacts
Resolutions
L1
Temporal Hierarchy by ω
Dynamical Generalizability Requirement
Dynamical Variables
L2
Overlooked Temporal Space $\mathbb{R}^T$
Amplified Inherent Biases
Relation-Indexed Representation
Neglected Dynamical Effects
Inverse Learning

Figure 1: Overview of the *Observation-Oriented* paradigm's primary limitations (labeled as $\boxed{\text{L1}}$ and $\boxed{\text{L2}}$). See section 1.2 for the concept of *hidden relation* ω, and 2.1 for *temporal space* $\mathbb{R}^T$ with relative timing axes.

This paper consists of four principal parts:

1. the Introduction: sets the foundation for the proposed *Relation-Oriented* perspective in section 1.1; also analyzes the roles of unobservable relational knowledge in modeling and uses an illustrative example to explain the resulting undetectable hierarchy (i.e., the limitation $\boxed{\text{L1}}$) in section 1.2.
2. Chapter I, including Sections 2 though 4: establishes the *Relation-Oriented* framework, to precisely decompose relationship modeling from a distributional perspective; and through which, to examine the fundamental impacts of the outlined limitations, and addresses the queries listed above.
3. Chapter II, from Sections 5 to 7: introduces the *Relation-Indexed Representation Learning* (RIRL) methodology as a baseline realization of the *Relation-Oriented* paradigm and evaluates the efficacy of using relation-indexed (i.e., causal) representations to identify dynamical effects.
4. the Conclusion in Section 8: summarizes the insights and findings of this study.

1.1 Relation-Oriented Perspective

Typically, experiments with n trials produce instances $x^n = x_1, \ldots, x_n$ from sequential random variables $X^n = X_1, \ldots, X_n$, which are usually assumed to be independent and identically distributed (i.i.d). When they evolve over time, n is often replaced by the timestamp t to get a temporal sequence $X^t = X_1, \ldots, X_t$, maintaining the i.i.d assumption, and the relationship function is usually in shape $Y = f(X^t; \theta)$.

In this study, we abandon the assumed independence over $\{X_i \mid i = 1, \ldots, t\}$ on the temporal dimension $\mathbf{t}$, instead treat their sequence X^t as a single entity, denoted by variable $\mathcal{X} \in \mathbb{R}^{d+1}$, with d representing the observational dimension of each instance X_i. For clarity, we use $X \in \mathbb{R}^d$ to represent a solely observational variable, and let $\mathcal{X} = \langle X, \mathbf{t} \rangle \in \mathbb{R}^{d+1}$ derived by incorporating the $\mathbf{t}$-dimension to encompass features across both observational and temporal dimensions. It is worth noting that variables such as $\mathcal{X}$ are conventionally referred to as spatial-temporal Andrienko (2003), while in this context, "spatial" is broadly interpreted to mean "observational", not restricted to be physically spatial like the geographic coordinates.

Consider the functional relationship model $\mathcal{Y} = f(\mathcal{X}; \theta)$, where $\mathcal{Y} = \langle Y, \tau \rangle \in \mathbb{R}^{b+1}$ with τ representing the temporal evolution of $Y \in \mathbb{R}^b$. We employ the Fisher Information $\mathcal{I}_{\mathcal{X}}(\theta)$ Ly et al. (2017) of $\mathcal{X}$ about θ, to define the component of $\mathcal{Y}$ (signified as $\hat{\mathcal{Y}}$) that is sufficiently identified by indexing through θ:

Definition 1. the *Relation-Indexed Representation* $\hat{\mathcal{Y}}_\theta$ in Relationship Modeling.

Let the ***relation*** θ adequately represents the influence of $\mathcal{X}$ on $\mathcal{Y}$, denoted as $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$, then $\hat{\mathcal{Y}}_\theta = f(\mathcal{X}; \theta)$ represents the *sufficient* component of $\mathcal{Y}$ about θ, which is, $\mathcal{I}_{\hat{\mathcal{Y}}_\theta}(\theta) = \max\ \mathcal{I}_{\hat{\mathcal{Y}}}(\theta) = \mathcal{I}_{\mathcal{X}}(\theta)$.

Consequently, $\hat{\mathcal{Y}}_\theta$ encapsulates the information within $\mathcal{Y}$ that is entirely derived from $\mathcal{X}$, thus defined as the *relation-indexed representation*. Accordingly, the remaining component of $\mathcal{Y}$, expressed as $\mathcal{Y} - \hat{\mathcal{Y}}_\theta$, does not depend on θ. The proposed *Relation-Oriented* modeling focuses on realizing the indexing role of θ.

The notation "→" typically denotes causality, although a directional relationship does not necessarily imply causation in logic. Nonetheless, for clarity, we will adopt terminology consistent with causal inference: for relationship $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$, we refer to $\mathcal{X}$ as the *cause* and $\mathcal{Y}$ as the *effect*, with a *relation* θ connecting them. Accordingly, the defined $\hat{\mathcal{Y}}_\theta$ aligns with the "causal representation" concept Scholkopf (2021). Crucially, through this paper, both *causality* and *correlation* denote types of relationships with a relation θ (their difference will be discussed later), while *association* refers to statistical dependency (typically nonlinear) between entities without an informative θ, expressed as $(\mathcal{X}, \mathcal{Y})$.

Remark 1. Given $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$ with ***observables*** $\mathcal{X}$ and $\mathcal{Y}$, the relationship model $\mathcal{Y} = f(\mathcal{X}; \theta)$ becomes *informative* due to the ***unobservable*** θ.

The outlined Remark 1 has its origins in the principle of Common Cause Dawid (1979); Scholkopf (2021), suggesting that any nontrivial (i.e., informative) conditional independence between two observables requires a third, mutual cause (i.e., the unobservable "relation" in our context).

$\mathcal{X}$ and $\mathcal{Y}$ can be either solely observational entities, equal to X and Y (e.g., images, spatial coordinates of a quadrotor, etc.), or observational-temporal entities (e.g., trends of stocks, a quadrotor's trajectory, etc.). Regardless of their characterization, the primary goal of adopting $\mathcal{Y} = f(\mathcal{X}; \theta)$ is to encapsulate the unobservable relational knowledge represented by θ, rather than merely distributional association $(\mathcal{X}, \mathcal{Y})$.

To clarify the concept of informative θ, consider a simple example: In the relationship "Bob (represented as X) has a son named Jim (represented as Y)", the father-son relational information $\mathcal{I}_{\mathcal{X}\mathcal{Y}}(\theta)$ between them is evident to humans, but unobservable to AI systems provided sufficiently observed social activities. Also, θ can be seen as the common cause of X and Y that makes their connection unique, rather than any random pairing of "Bob" and "Jim". Through the provided observations, AI may deduce a particular associative pattern over (X, Y), but cannot internalize the unobservable information $\mathcal{I}_{\mathcal{X}\mathcal{Y}}(\theta)$.

Based on the symbolization in Definition 1 and the principle of Remark 1, a *Relation-Oriented* framework will be established in Section 2 to offer more complete insides into relationship modeling.

1.2 Unobservable Relational Knowledge

The unobservable relations in knowledge may not directly serve as the learning objective θ, but still be relative to and profoundly impact the modeling process. We elucidate this phenomenon through an example: Notably, on social media, AI-created personas can have realistic faces but seldom showcase hands. This is because AI for visual tasks struggles with the intricate structure of hands, instead treating them as arbitrary assortments of finger-like items. Figure 2(a) provides 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***, guiding us to visit specific mental representations Pitt (2022). As illustrated in (b), our cognitive process operates hierarchically, through a series of relations, denoted by $\theta = \{\theta_i, \theta_{ii}, \theta_{iii}\}$. 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.

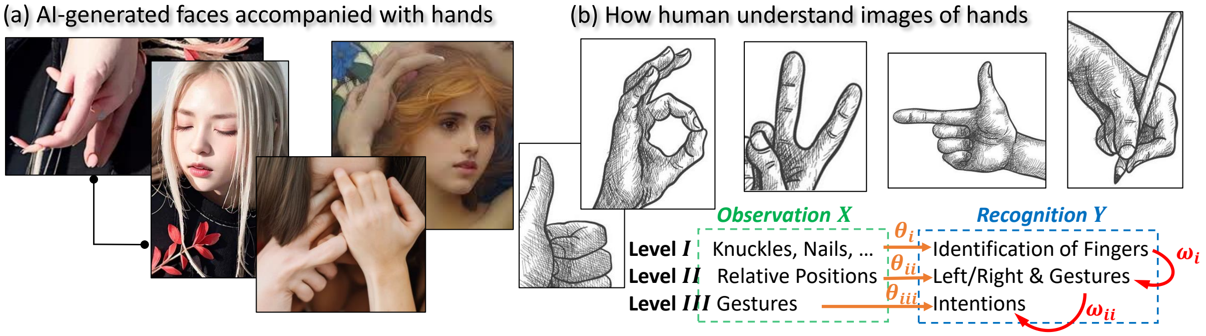


Figure 2: Unobservable relations $\theta = \{\theta_i, \theta_{ii}, \theta_{iii}\}$ and $\omega = \{\omega_i, \omega_{ii}\}$. AI can generate reasonable faces but treat hands as arbitrary mixtures of fingers; while human cognition processes observations hierarchically to avoid this mess, indexing through a series of relations $\{\theta_i, \theta_{ii}, \theta_{iii}\}$.

Typically, these visual learning tasks do not directly capture relational information, neither $\mathcal{I}_X(\theta)$ nor $\mathcal{I}_X(\omega)$, focusing instead on modeling the entity Y solely based on observations X. Without indexing through θ, AI systems may struggle to distinguish different levels in Y. They tend to encapsulate the observational dependences, such as $(X_{II} \mid X_I)$ and $(X_{III} \mid X_I, X_{II})$, entirely within the association (X_I, X_{II}, X_{III}), resulting a lack of informative insights into ω.

However, the hidden ω may not always be essential. For instance, AI can generate convincing faces because the appearance of eyes θ_i strongly indicates the facial angles θ_{ii}, i.e., $\mathcal{I}_X(\theta_{ii}) \not> \mathcal{I}_X(\theta_i)$, removing the need to distinguish eyes Y_I from faces Y_{II} to reveal $\mathcal{I}_X(\omega_i) = 0$. Furthermore, with observational levels in X fully captured, AI can inversely detect indexing relations using methods like reinforcement learning Sutton (2018); Arora (2021). For example, in Figure 2, when AI systems receive approval for generated five-fingered hands $(Y_{II} \mid Y_I)$, reflecting ω_i through $(X_{II} \mid X_I)$, they may autonomously begin to derive $X_I \xrightarrow{\theta_i} Y_I$.

Definition 2. <u>*Hidden Relation*</u> ω and its resulting <u>*Knowledge Hierarchy*</u>.
Unlike the ***indexing relation*** θ as a learning objective, the ***hidden relation*** ω forms hierarchical knowledge levels, necessitating model *generalizability* to maintain effectiveness across.

The illustration in Figure 2 shows different roles of unobservable relations, θ and ω, in solely observational learning tasks. Our main focus, however, is on relationship models that explicitly incorporate θ as a functional parameter. A *generalizable* model allows lower-level knowledge like θ_i to be reusable for higher-level learning tasks like θ_{ii} Scholkopf (2021), reflecting our innate ability to generalize knowledge cognitively. Generalizable also denotes the capacity to *individualize* from higher to lower levels, to accommodate different ω_i values.

Consider this example: Family incomes X influence grocery shopping frequencies Y through relation θ. Here, the cultural background ω emerges as an important factor, such that an effective model $Y = f(X;\theta)$ has to be individualizable, i.e., conditioned on a specific country (represented by a particular ω value) to ensure practical utility. On the opposite, a generalization would imply $\omega = \varnothing$.

For the sake of clarity, hereafter in this paper, unless explicitly stated otherwise, the hidden relation ω represents two hierarchical levels: The generalized level $X_o \xrightarrow{\theta_o} Y_o$ with θ_o implying $\omega = \varnothing$, and the individualized level $X_\omega \xrightarrow{\theta_\omega} Y_\omega$ given θ_o with a specific ω value, collectively notated as $(\theta,\omega) = \begin{pmatrix}\theta_o \\ \theta_\omega\end{pmatrix}$.

In the context of *causal* relationship learning, the temporal events for $\mathcal{X}$ and $\mathcal{Y}$ are usually pre-identified. However, this may not guarantee their *temporal features* to be completely captured by the model, making (θ,ω) ***undetectable*** for AI models, and precluding methods like inverse reinforcement learning.

Chapter I: Limitations of Current Observation-Oriented Paradigm

The prevailing *Observation-Oriented* machine learning paradigm misaligns with the relation-centric essence of human comprehension Pitt (2022), which may not have been critical in the past. In traditional causal inference, challenges could be addressed through intended adjustments due to the limited scale of questions. Nonetheless, with the advancements in AI-based large models, the consequences of this misalignment have become increasingly significant across various applications.

Section 2 introduces a *Relation-Oriented* dimensionality framework, representing relationships as decomposed distributions. This framework underscores the critical role of relative timings in structural causal reasoning (highlighted as limitation $\boxed{\text{L2}}$), and the importance of dynamic capturing for generalizable causal models. Subsequently, Section 3 explores the implications of often-overlooked dynamical effects (the secondary impact of $\boxed{\text{L2}}$), mainly in response to the ❖ outlined challenges. Lastly, Section 4 thoroughly examines the scheme and impact of inherent biases due to overlooking the underlying relative timings in structural causal models (the primary impact of $\boxed{\text{L2}}$).

2 Relation-Oriented Dimensionality Framework

A central question in the debates surrounding AGI persists: Can AI systems, which rely on mathematical symbolizations, achieve a human-like understanding sufficient to handle empirical inquiries Newell (2007); Pavlick (2023)? We propose to focus on representing unobservable elements within our knowledge, such as abstractly meaningful relations, which are vital for the informativeness of causal reasoning. By indexing through these relations, AI models have the potential to reflect our logical deductions, embody the cognitive concepts they lead to, and ultimately construct their representations. As aligning with our causal knowledge, these representations can yield generalizable models, critical for actualizing causal reasoning in AGI.

By Definitions 1 and 2, representing a relationship necessitates two types of variables: the observables $\{\mathcal{X},\mathcal{Y}\}$, and the unobservables (θ,ω). As specified, $\mathcal{X}$ and $\mathcal{Y}$ include both *observational* and *temporal* features. In response, we adopt the concept of a *hyper-dimension* to integrate these unobservable features. Consequently, we establish a framework, as illustrated in Figure 3, to represent relationships as joint distributions across three distinct types of dimensions. For clarity, “feature” refers to the potential variable fully representing a certain distribution of interest.

Figure 3 aims to decompose our cognitive space where relational knowledge is stored. The hyper-dimensional space $\mathbb{R}^H$ is constructed by aggregating all ***unobservable*** relations in our knowledge, such as $(\theta,\omega) \in \mathbb{R}^H$. Conversely, the observational-temporal joint space, $\mathbb{R}^O \cup \mathbb{R}^T$, is considered as the ***observable*** space. In both $\mathbb{R}^O$ and $\mathbb{R}^T$, a temporal dimension consistently signifies the evolution of timing but represents distinct concepts, as outlined in section 2.1. Within such a dimension, linear and nonlinear distributions correspond to *static* and *dynamical* features, respectively, a distinction further explained in section 2.2.

Definition 3. The *Relationship Representation* within the proposed Dimensionality Framework.

For the relationship $\mathcal{X} \xrightarrow{\vartheta} \mathcal{Y}$, where $\{\mathcal{X}, \mathcal{Y}\} \in \mathbb{R}^O$, the ***structuralized relation*** $\vartheta \in \mathbb{R}^T \cup \mathbb{R}^H$ can be decomposed as:

$$\vartheta = \overrightarrow{\theta^1 \dots \theta^T}, \text{ where } (\theta^i, \theta^j) \in \mathbb{R}^H \text{ for any } i \neq j \in \{1, \dots, T\}. \text{ Accordingly,}$$

$$(\vartheta, \omega) = \begin{pmatrix} \vartheta_o \\ \vartheta_\omega \end{pmatrix} = \begin{pmatrix} \theta_o^1 & \dots & \theta_o^T \\ \theta_\omega^1 & \dots & \theta_\omega^T \end{pmatrix} \text{ with any } (\theta_o^i, \theta_o^j) \in \mathbb{R}^H \text{ and } (\theta_\omega^i, \theta_\omega^j) \in \mathbb{R}^H.$$

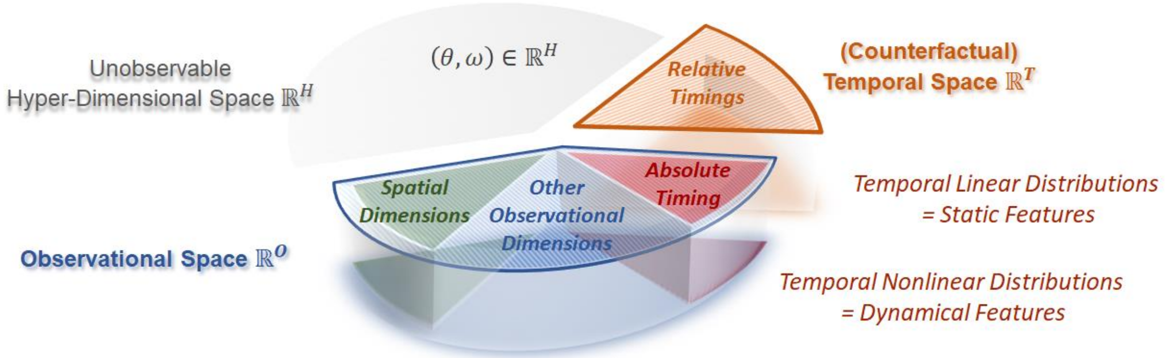

Unobservable
Hyper-Dimensional Space $\mathbb{R}^H$
$(\theta, \omega) \in \mathbb{R}^H$
Relative Timings
(Counterfactual)
Temporal Space $\mathbb{R}^T$
Spatial Dimensions
Other Observational Dimensions
Absolute Timing
Temporal Linear Distributions
= Static Features
Observational Space $\mathbb{R}^O$
Temporal Nonlinear Distributions
= Dynamical Features

Figure 3: *Relation-Oriented* Dimensionality Framework: splitting the knowledge-storing cognitive space by their accommodated features, where $\{X^t, Y^t\} \in \mathbb{R}^{O-1}$, $\{\mathcal{X}, \mathcal{Y}\} \in \mathbb{R}^O$, and $(\mathcal{X}, \mathcal{Y} \mid \vartheta) \in \mathbb{R}^O \cup \mathbb{R}^T$.

2.1 Absolute Timing vs. Relative Timings

In spatial-temporal data, the attribute recording observed timestamp t typically reflects the ***absolute*** timing of reality. However, from a modeling view, the temporally meaningful t values are indistinguishable from other attributes. As shown in Figure 3, the absolute timing $\mathbf{t}$ serves as a standard dimension within the observational space $\mathbb{R}^O$, along which, $\mathcal{X}$ and $\mathcal{Y}$ are invariably observed as data sequences X^t and Y^t.

Contrarily, in our cognition, ***relative*** timings inherently exist Wulf (1994) to support the "what if" thinking and form structualized causal knowledge. We thereby designate a distinct "temporal space" $\mathbb{R}^T$, composed of T relative timings as axes (i.e., T cognitive *timelines* Shea (2001)), to accommodate the knowledge-aligned (i.e., under ϑ, as per Definition 3) temporal distributions. Instead of as $\mathcal{Y} \in \mathbb{R}^O$, it can distribute across $\mathbb{R}^T$, represented as $(\mathcal{Y} \mid \mathcal{X}, \vartheta) \in \mathbb{R}^{O-1} \cup \mathbb{R}^T$, or jointly represented as $(\mathcal{X}, \mathcal{Y} \mid \vartheta) \in \mathbb{R}^O \cup \mathbb{R}^T$.

ϑ can span up to T timing dimensions in $\mathbb{R}^T$, with the effect $\mathcal{Y} = \sum_{i=1}^T \hat{\mathcal{Y}}^i$ decomposed into T components, each residing in a distinct timing. Crucially, defining ϑ as a "structuralized" relation not only recognizes its multi-dimensionality but also highlights the potential *nonlinear dependence* among these timings, manifested as $(\hat{\mathcal{Y}}^i, \hat{\mathcal{Y}}^j) \in \mathbb{R}^{O-1} \cup \mathbb{R}^T$, while more precisely represented by $(\theta^i, \theta^j) \in \mathbb{R}^H$ in Definition 3. We term these nonlinear temporal dependences as ***dynamical interactions*** for clarity, which necessitate the establishment of a distinct $\mathbb{R}^T$ space, rather than additional temporal dimensions within $\mathbb{R}^O$ (detailed in section 2.3).

For instance, patients' vital signs are recorded daily in a hospital with *absolute* chronological timestamps. However, to assess a medical intervention $\mathcal{Y}$, a uniform series of post-medication events must be selected, for example, spanning from the day after medication to the 30th day. This creates a timeline represented by the axis ticked as $[1, 30]$ to denote the *relative* timing, regardless of *absolute* timestamps of the selected records. Yet, if the intervention involves two distinct aspects, such as the primary effect $\hat{\mathcal{Y}}^1$ and the side effect $\hat{\mathcal{Y}}^2$, and their mutual influences are of interest, then two separate relative timings, $\mathbf{t_1}$ and $\mathbf{t_2}$, must be considered for their individual evolutions, even though both may be labeled as $[1, 30]$.

Remark 2. Although $\mathcal{Y} \in \mathbb{R}^O$ is *observed* as a sequence along the absolute timing $\mathbf{t}$, it may represent an *underlying* structure determined by $\mathcal{X} \xrightarrow{\vartheta} \mathcal{Y}$, spinning multiple relative timing axes in $\mathbb{R}^T$ space.

Conventionally, the concept of "temporal dimension" is often simplified to be the single absolute timing $\mathbf{t}$, evident from the traditional "spatial-temporal" analysis Alkon (1988); Turner (1990); Andrienko (2003), to recent advancements in language models Wes (2023). However, as emphasized in Remark 2, our cognitive perception of "time" is more complex, fundamentally enabling our causal reasoning Coulson (2009).

For an intuitive insight into the implications of neglecting relative timings in $\mathbb{R}^T$, let's consider an analogy: Imagine ants dwelling on a floor's two-dimensional plane. To predict risks, the scientists among them create two-dimensional models and instinctively adopt the nearest tree as a height reference. They noticed increased disruptions at the tree's first branch, which indeed correlates to the children's heights, given their curiosity. However, without understanding humans as three-dimensional beings, they can only interpret it by adhering to the first branch. One day, after relocating to another tree with a lower height, the ants found the risk presenting at the second branch instead, making their model ineffective. They may conclude that human behaviors are too complex, highlighting the model generalizability issue.

As three-dimensional beings, we inherently lack the capacity to fully integrate the fourth dimension - time - into visual perception. Instead, we conceptualize "space" in three dimensions to incorporate features of the temporal dimension along a *timeline* within the space, analogous to our "tree". Yet, ants do not need to fully comprehend the three-dimensional world to build a generalizable model; instead, they need only recognize the "forest" out of their vision (i.e., counterfactual), which consists of all "possible trees" with *relatively* different branch locations. Similarly, in our modeling, we must include the $\mathbb{R}^T$ space, composed of all potential relative timings within our causal knowledge, although they cannot be directly observed.

Remark 3. *Counterfactuals* can be considered as posterior distributions within $\mathbb{R}^O \cup \mathbb{R}^T$.

Addressing the counterfactual query "what effect would be if the cause were changed" differentiates causality from mere correlations Scholkopf (2021). In the proposed framework, counterfactuals are more intuitively interpreted through distributions, potentially offering valuable insights in fields like quantum computing. Specifically, the observed prior conditions $\mathcal{X}$ can be considered as features in $\mathbb{R}^O$, whose effects $\mathcal{Y}$ act as a conditional distribution within $\mathbb{R}^{O-1} \cup \mathbb{R}^T$, incorporating T possible observed timings in the future.

2.2 Dynamical vs. Sequential Static

The distributions along a dimension can be broadly classified into *linear* and *nonlinear* categories. Within the temporal dimension, these correspond to ***static*** and ***dynamical*** temporal features, respectively, and can be represented by corresponding variables. Static features are typically linked to specific timestamps. For instance, consider the statement "rain leads to wet floors"; here "wet floors" represents a state that can be identified at a particular point in time. Therefore, it can be denoted as a static variable X_t with a specified timestamp t. In contrast, the expression "floors becoming progressively wetter" necessitates a representation that captures the temporal distribution, to account for changes over time, like $X^t = X_1, \ldots, X_t$. However, this raises a question: Is X^t a dynamical variable or a sequence of static variables?

Within the current machine learning paradigm, the distinction between "static" and "dynamical" is typically made between "models" instead of "variables" PGMadhavan (2016), which refers to whether time is a factor in the model's equations. However, this essentially requires the function $f(X^t; \theta)$ to represent the *dynamics of effect* inherently encompassed by $\mathcal{Y}$. As a result, the model assumption for $f(; \theta)$, as well as the identification of a *static* outcome Y_{t+1}, become crucial in determining how much effect dynamics can be captured Weinberger & Allen (2022), or potentially neglected, which will be discussed further in Section 3.

Definition 4. The *Dynamically Significant* $\mathcal{Y}$ vs. Sequential Static Y^τ.
As a *dynamical* variable, $\mathcal{Y} = \langle Y, \tau \rangle \in \mathbb{R}^O$ permits ***nonlinear*** *computational freedom* over τ, whereas a *sequential static* variable $Y^\tau \in \mathbb{R}^{O-1}$ assumes i.i.d or ***linear*** changes along τ. They samely manifest as sequential instances $y^\tau = y_1, \ldots, y_\tau$, while the dynamical significance of $\mathcal{Y}$ is ***model-dependent***.

Definition 4 is based on the proposed *Relation-Oriented* paradigm: The relation $\theta \in \mathbb{R}^H$ and the outcome $\mathcal{Y} \in \mathbb{R}^O$ are considered individually, where θ represents certain unobservable information within $\mathbb{R}^H$, lacking an explicit distributional representation. This allows $\mathcal{Y}$ to be considered a variable that encompasses the dynamical effects caused by $\mathcal{X}$. Similarly, the cause $\mathcal{X} = \langle X, t \rangle \in \mathbb{R}^O$ can also be a dynamical variable whose fulfillment depends on specific models. For example, RNN models typically formulate $Y_{t+1} = f(\mathcal{X}; \theta)$ with a dynamical cause represented by latent space features, but remaining the outcome static.

Accordingly, the statement "floors becoming progressively wetter" can be roughly considered as "linearly increasing from 0% to 100% in 10 minutes" to be a sequential static feature. It can also be depicted as a continuous nonlinear distribution, a dynamical feature for finer granularity. The latter can cover variances in the former, such as varying progression speeds, which the former cannot. In essence, the fulfillment of dynamical variables is crucial for achieving model generalizability over temporal dimensions.

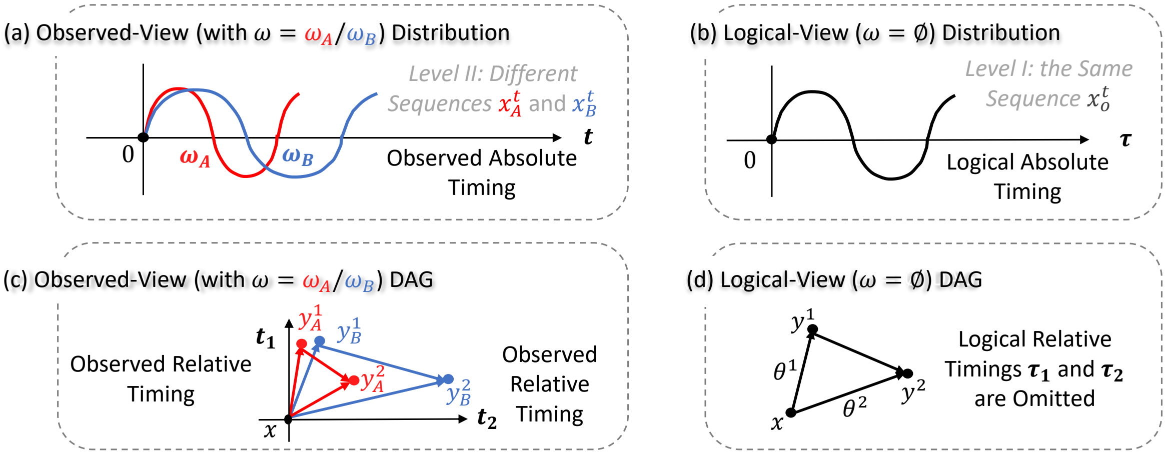

(a) Observed-View (with $\omega = \omega_A/\omega_B$) Distribution
Level II: Different Sequences x_A^t and x_B^t
t
0
ω_A
ω_B
Observed Absolute Timing
(b) Logical-View ($\omega = \emptyset$) Distribution
Level I: the Same Sequence x_o^t
τ
0
Logical Absolute Timing
(c) Observed-View (with $\omega = \omega_A/\omega_B$) DAG
Observed Relative Timing
t_1
y_A^1
y_B^1
y_A^2
y_B^2
x
t_2
Observed Relative Timing
(d) Logical-View ($\omega = \emptyset$) DAG
y^1
θ^1
x
θ^2
y^2
Logical Relative Timings τ_1 and τ_2 are Omitted

Figure 4: Comparisons of the individualized dynamics from the model's Observed-View, and the generalized dynamics from humans' Logical-View. In (c) and (d), the structuralized relation $\vartheta = \overrightarrow{\theta^1\theta^2}$ with $T = 2$.

Considering a structuralized relation ϑ, a valid model-generalizing process requires the model to remain effective over temporal dimensions at any level, including both the absolute timing within $\mathbb{R}^O$ and the relative timings in counterfactual $\mathbb{R}^T$. In our cognition, the generalized causal knowledge ($\omega = \varnothing$) can be instinctively extracted from individualized varied scenarios (with varying ω values). However, the undetectability of (ϑ, ω) implies our models cannot autonomously fulfill this process, irrespective of whether they are AI-based.

Figure 4 showcases models' and humans' perspectives, distinguished as the "Observed-View" and "Logical-View". (a) and (b) compare a dynamical distribution along absolute timing in $\mathbb{R}^O$, while (c) and (d) display a DAG structure across two relative timings in $\mathbb{R}^T$, which exhibits a typical *dynamical confounding* scenario.

In (c), the static instances y_A^1 and y_B^1 signify that the two individualized dynamical effects $\mathcal{Y}_A$ and $\mathcal{Y}_B$ reach the same status value y^1 in dimension $\mathbf{t_1}$, i.e., attaining an equivalent magnitude; this situation is similarly observed in another timing dimension $\mathbf{t_2}$. Notably, the edge from y^1 to y^2 indicates an interaction between the two dynamical effect components $\hat{\mathcal{Y}}^1$ and $\hat{\mathcal{Y}}^2$, which can be either static or dynamical, suggesting their linear or nonlinear dependence (detailed definitions are provided in Section 4).

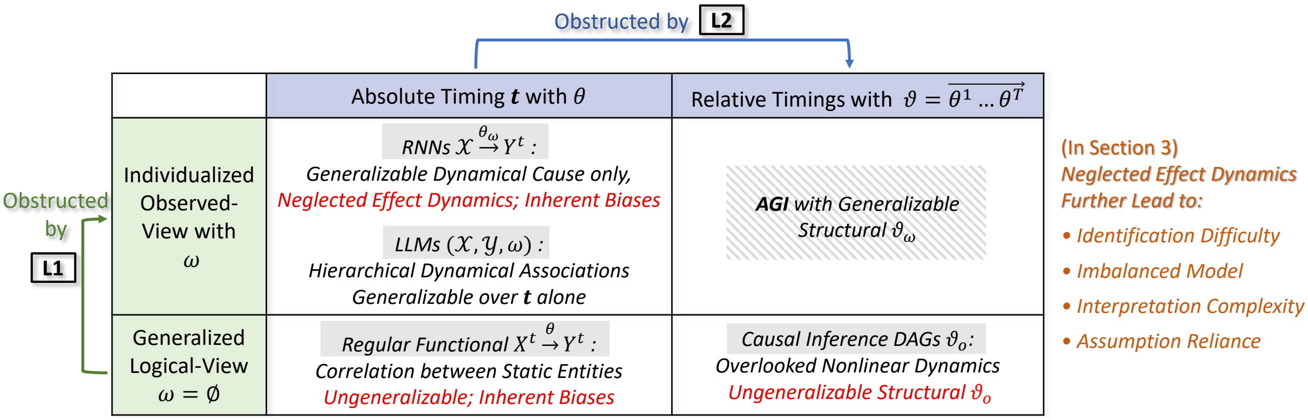
Definition 5. The <u>*Dynamical Confounding*</u> Phenomenon.

For relationship $\mathcal{X} \xrightarrow{\vartheta} \mathcal{Y}$, when dynamical effect $\mathcal{Y}$ comprises multiple components over distinct relative timings, the *interaction* among them can lead to *dynamical confounding* within $\mathbb{R}^{O-1} \cup \mathbb{R}^T$.

2.3 Informative Hyper-Dimensional Space

In summary, our structural causal reasoning can be represented as $(\vartheta, \omega) \in \mathbb{R}^T \cup \mathbb{R}^H$. Accordingly, AGI that meets our expectations should adequately encapsulate informative ϑ and ω. Here, $\vartheta \in \mathbb{R}^T \cup \mathbb{R}^H$ denotes the structuralized causality within our knowledge, while $\omega \in \mathbb{R}^H$ indicates the ability to capture nonlinearities in all dimensions (including temporal dynamics), to achieve model generalizability.

Figure 5 provides a fundamental overview of prevailing relationship models, highlighting their intrinsic limitations, as outlined in Figure 1. In this context, ϑ_ω is used to represent generalizable causal structures within AGI, and we accordingly summarize the two major obstacles in our pursuit of it.


Obstructed by L2
Obstructed by L1
(In Section 3)
Neglected Effect Dynamics Further Lead to:
• Identification Difficulty
• Imbalanced Model
• Interpretation Complexity
• Assumption Reliance

	Absolute Timing $\boldsymbol{t}$ with θ	Relative Timings with $\vartheta = \overrightarrow{\theta^1 \ldots \theta^T}$
Individualized Observed-View with ω	RNNs $\mathcal{X} \xrightarrow{\theta_\omega} Y^t$: *Generalizable Dynamical Cause only,* *Neglected Effect Dynamics; Inherent Biases* LLMs $(\mathcal{X}, \mathcal{Y}, \omega)$: *Hierarchical Dynamical Associations* *Generalizable over* $\boldsymbol{t}$ *alone*	***AGI*** *with Generalizable Structural* ϑ_ω
Generalized Logical-View $\omega = \emptyset$	*Regular Functional* $X^t \xrightarrow{\theta} Y^t$: *Correlation between Static Entities* *Ungeneralizable; Inherent Biases*	*Causal Inference DAGs* ϑ_o: *Overlooked Nonlinear Dynamics* *Ungeneralizable Structural* ϑ_o

Figure 5: Overview of major obstacles toward AGI (seeing Figure 1): $\boxed{\text{L1}}$ = Undetectable temporal hierarchy ω requiring dynamical generalizability. $\boxed{\text{L2}}$ = Overlooked multi-dimensional $\mathbb{R}^T$ comprising relative timings.

Regular relationship models derive the functional parameter θ from the correlation between static cause and effect events, X^t and Y^t, priorly identified using absolute timestamps. Notably, Granger causality Granger (1993), a method well-regarded in economics Maziarz (2015), introduces separate temporal sequences for cause X^t and effect Y^τ, suggesting multiple timings. However, without nonlinear computational capabilities over them, distinguishing between $\mathbf{t}$ and τ for static timestamps offers limited meaning.

Remark 4. The significance of *Temporal Dimensions* lies in allowing distinct dynamical evolutions.

As depicted in Figure 4 (d), causal inference often omits explicit relative timing axes in causal DAGs, because of the typical exclusion of nonlinear dynamics. While inherently adopting a *Relation-Oriented* perspective based on the Logical-View structural knowledge ϑ_o, it tends to overlook the Observed-View scenarios with varied ω, thus failing to exhibit causal models' *dynamical generalization* needs. To address this, we suggest enhancing DAGs to visualize dynamical variations across multiple timings, as introduced in Section 4.

AI-based RNNs are increasingly favored in modern relationship learning Xu et al. (2020), a trend that reflects their proficiency in handling temporally nonlinear causes. RNNs transform the sequence X^t into a feature representation in the latent space, enabling nonlinear computation over $\mathbf{t}$ to effectively fulfill dynamical $\mathcal{X}$. However, potential dynamics of the effect $\mathcal{Y}$ are often overlooked, resulting in an *imbalanced* causal model function $Y_{t+1} = f(\mathcal{X}; \theta)$ with a static outcome Y_{t+1}. This accordingly motivates the emerging trend in *inverse learning* methods Arora (2021). Further details will be discussed in Section 3.

Notably, large language models (LLMs) are able to identify different temporal changes in the semantic space, including nonlinear ones Wes (2023). However, "multiple temporal dimensions" to accommodate distinct dynamics do not necessarily equate to "multiple relative timings".

Remark 5. *Temporal Dimensions* with ***nonlinear independence*** can be simultaneously identified from absolute timing $\mathbf{t}$ within $\mathbb{R}^O$; while *Relative Timings* indicate potential ***nonlinear dependence***, i.e., dynamical interactions, requiring a counterfactual space $\mathbb{R}^T$ to house the underlying structure.

Current LLMs primarily focus on capturing semantic associations based on the absolute timing $\mathbf{t}$, which indicates the order of phrases. This approach categorizes them within the scope of solely observational learning, lacking a *Relation-Oriented* perspective. However, considering the consistent sequential semantics in words, overlooking relative timings is reasonably justifiable for basic context-associative learning needs.

Nevertheless, the association $(\mathcal{X}, \mathcal{Y})$ captured along $\mathbf{t}$ only reflects θ rather than explicitly representing it, meaning it does not extract $\mathcal{I}(\theta)$. This may contribute to AI's ability to generate intelligent responses without truly "understanding" in the human sense, due to the lack of an informatively represented θ.

Given the adaptability of meta-learning to diverse observational learning tasks Hospedales et al. (2021), its integration with LLMs could significantly improve the generalizability of associative models Lake (2023). This application may enable capturing $(\mathcal{X}, \mathcal{Y}, \omega)$ to reflect hierarchical relations (θ, ω) over the timing $\mathbf{t}$. However, compared to our goal of explicitly representing ϑ_ω, with encapsulated $(\theta^i, \theta^j) \in \mathbb{R}^H$ for all $i \neq j \in \{1, \dots, T\}$ to achieve causal reasoning, discussing AGI within the current LLM framework may still be premature. We suggest enabling *Relation-Oriented* meta-learning could be a significant step towards this goal.

3 Neglected Dynamical Effects in Causal Learning

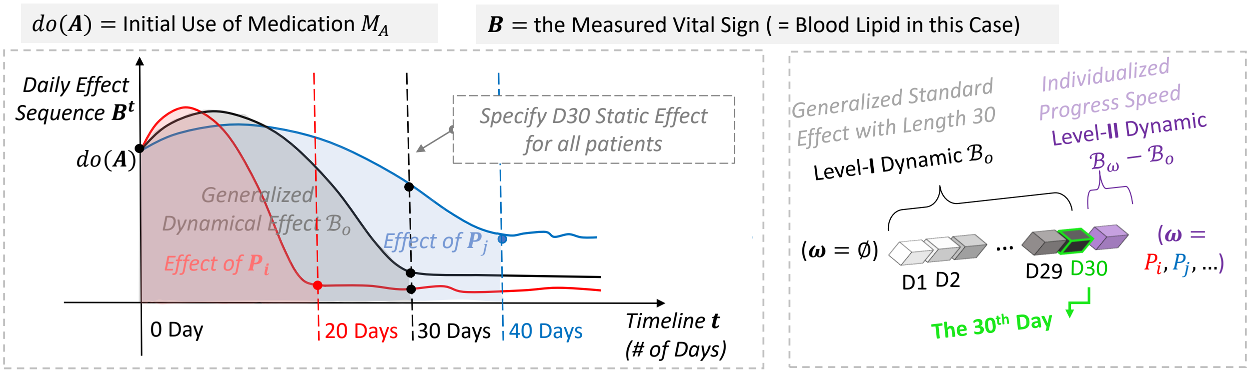
Traditional causal inference emphasizes the interpretability of causal models, particularly in differentiating them from mere correlations. These distinctions, while not inherently integrated into the modeling context, are mainly evident in model interpretations. Despite the statistical basis of causal inference, the importance of nonlinear temporal dynamics is yet to be fully acknowledged. This section focuses on these frequently overlooked dynamics, striving to provide a more intuitive understanding of causal learning.

Definition 6. Causality vs. Correlation in the modeling context.

- Causality $\mathcal{X} \xrightarrow{\vartheta} \mathcal{Y}$ is the relationship neccessitating dynamical effect $\mathcal{Y} \in \mathbb{R}^{O-1} \cup \mathbb{R}^T$.
- Correlation $X^t \xrightarrow{\theta} Y^t$ only requires static cause and effect, possibly sequential X^t and Y^t.

The timestamp t, first introduced by the Picard-Lindelof theorem in the 1890s, initiates the functional form $Y_{t+1} = f(X_t)$ to represent time evolution. Then, time series learning methods, like autoregressive models Hyvärinen (2010), facilitate the form of $Y_{t+1} = f(X^t)$ using a sequential causal variable X^t with a predetermined time progress from t to $t + 1$.

For RNNs, the latent space optimization over X^t is driven by predicting observed Y_{t+1} value through the parameterized relation θ, enabling the form of $Y_{t+1} = f(\mathcal{X}; \theta)$ with a dynamical cause $\mathcal{X}$. However, the effect Y_{t+1} remains static, with its potential dynamics governed by the function f. While f can be selected as linear or nonliner, it influences $\mathcal{X}$ only but still leave the time evolution from t to $t + 1$ as ***linear***.



(a) Observed Temporal Distribution $\mathcal{B}_\omega$ (b) Complete Two-Level Effect Dynamics

Figure 6: $do(A)$ denotes the initial use of medication M_A for reducing blood lipid B. The goal is to estimate the generalized effect of M_A, i.e., $\mathcal{B}_o$. By the rule of thumb, $\mathcal{B}_o$ 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.

The example in Figure 6 illustrates the often overlooked effect dynamics in traditional causal models. The action $do(A)$ causes dynamical $\mathcal{B}_\omega$ (observed as sequence B^t), disentangled by two levels in (b): Level **I**, the generalized standard sequence $\mathcal{B}_o$ of length 30; Level **II**, the individualized variations $\mathcal{B}_\omega - \mathcal{B}_o$. Assume the unobserved individualized characteristics linearly impact $\mathcal{B}_o$, making $\omega = P_i, P_j, \ldots$ simply represent speeds.

A typical clinical model, like $B_{t+30} = f(do(A_t))$ that averages all patients' D30 static effects as the outcome, turns to neglect D1-D29 within $\mathcal{B}_o$. However, even adopting a sequential outcome B^t (e.g., Granger causality), it remains challenging to accurately estimate $\mathcal{B}_o$ by linear averaging, not to mention further reaching $\mathcal{B}_\omega$. Particularly, it requires the selected records to meet certain criteria, essentially equal to manually defining the boundary of $\mathcal{B}_o$ by exploring all possible ω values.

Such hierarchical dynamical effects are prevalent in fields like epidemic progression, economic fluctuations, strategic decision-making, etc. They often rely on a similar strategy to manually identify specific levels, e.g., the group-specific learning methodology Fuller et al. (2007). These approaches have become impractical in AI-based applications and may lead to notable information loss in large-scale structural models.

3.1 Identification Difficulty of Nonlinear Effect

The *Observation-Oriented* modeling usually pre-identifies static effects based on existing knowledge, such as $Y_t = f(\mathcal{X};\theta)$, to derive θ and accordingly output static sequential estimations $\hat{Y_\theta^t} = \hat{Y}_1, \ldots, \hat{Y}_t$. Yet, two types of errors may challenge the accuracy: the discrepancy between the targeted nonlinear (i.e., dynamical) effect and the specified outcome sequence $\mid \mathcal{Y} - Y^t \mid$; and the approximation error due to the predetermined model function $f(;\theta)$. They contribute to the identification difficulty in causal learning Zhang (2012).

Specifically, due to static outcome, the burden of representing the dynamical aspects of $\mathcal{Y}$ shifts either to $f(;\theta)$ or to $\mathcal{X}$. In the former scenario, a factor σ signifying "disturbance" is integrated into the function, resulting in $f(;\theta+\sigma)$ Zhang (2012). In the latter case, as treated in do-calculus Pearl (2012); Huang (2012), the dynamical $\mathcal{X}$ needs to be manually discretized as temporal events to ensure their identifiable effects. This enables a fluid transformation from dynamical cause to observational effect, but the identifiability relies on non-experimental data (controllable θ) and can introduce additional complexities.

Considering the *differential* essence of do-calculus, we provide a streamlined reinterpretation of its three core rules from an *integral* viewpoint. Let $do(x_t) = (x_t, x_{t+1})$ indicate the occurrence of an instantaneous event $do(x)$ at time t, with the time step Δt appropriate to ensure the *interventional* effect of $do(x_t)$ identifiable as a function of the resultant distribution at $t+1$. Meanwhile, a separate *observational* effect is provoked by the static x_t value. Then, the dynamical cause $\mathcal{X}$ can be discretized as below:

$$\text{Given } \mathcal{X} \xrightarrow{\theta} Y, \text{ where } \mathcal{X} = \langle X, t\rangle \in \mathbb{R}^{d+1} \text{ with the augmented } \mathbf{t} \text{ dimension residing a } l\text{-length sequence,}$$

$$\mathcal{X} = \int_0^l do(x_t)\cdot x_t \; dt \quad \text{with} \begin{cases} (do(x_t)=1) \mid \theta, & \textit{Observational} \text{ only (Rule 1)} \\ (x_t = 1) \mid \theta, & \textit{Interventional} \text{ only (Rule 2)} \\ (do(x_t)=0) \mid \theta, & \text{No } \textit{interventional} \text{ (Rule 3)} \\ \text{otherwise} & \text{Associated } \textit{observational} \text{ and } \textit{interventional} \end{cases}$$

$$\text{The effect of } \mathcal{X} \text{ can be derived as } f(\mathcal{X}) = \int_0^l f_t\big(do(x_t)\cdot x_t\big)\; dt = \sum_{t=0}^{l-1}(y_{t+1}-y_t) = y_l - y_0$$

Based on a controllable θ, it addresses three criteria that can preserve conditional independence between *observational* and *interventional* effects, completing the chain rule, but sidesteps more generalized cases. If oppositely defining $\mathcal{Y} = \langle Y, \tau\rangle$ as a dynamical effect, discretizing it in $do(y)$ remains necessary.

3.2 Imbalance between Cause and Effect

For the modeling computation, causal directionality (i.e., the roles of cause and effect) may not impose restrictions, although it is often emphasized in model interpretations. Specifically, when selecting a model function for $X \to Y$, one could use $Y = f(X;\theta)$ to predict the effect Y, or $X = g(Y;\phi)$ to inversely infer the cause X. Both parameters, θ and ϕ, are obtained from the joint probability $\mathbf{P}(X,Y)$ without imposing modeling constraints. We refer to this as *symmetric directionality* for clarity.

Concern for causal model direction mainly arises for two reasons: 1) it aligns with our intuitive understanding of temporal progression, and 2) there is an inherent ***imbalance*** between how causal models capture dynamics of the cause and the effect, with RNNs as a typical example, formulated as $Y = f(\mathcal{X};\theta)$.

Given the symmetric directionality, inverse learning methods Arora (2021) capitalizing on this imbalance have recently garnered increasing attention, to achieve autonomous effect identification by inversely assigning the effect as the cause within RNNs. It is suitable for relationships along a single absolute timing, but not for addressing causal structures represented by ϑ. Specifically, the neglected relative timings implicitly assume nonlinear independence between distinct dynamical effects, which could lead to inherent bias regardless of the modeling direction. This will be further detailed in Section 4.

Another factor contributing to this imbalance is the increased empirical difficulty when specifying effect sequence Y^t compared to cause sequence X^t. While organizing time series data around a major causal event (e.g., days of heavy rain) is feasible, pinpointing the precise onset of subsequent effects (e.g., the exact day a flood began due to the rain) remains a more complex task.

Remark 6. By indexing through θ, the optimization of $\mathcal{X}$ and $\mathcal{Y}$ can be achieved simultaneously, mitigating their imbalance and enabling autonomous effect identification.

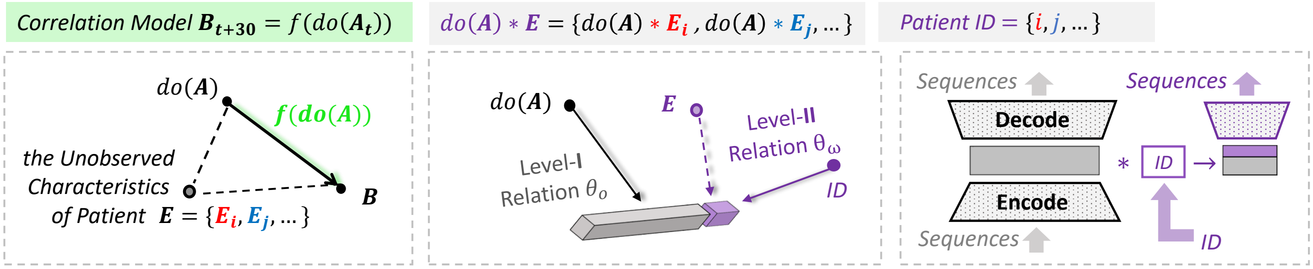
The proposed *Relation-Oriented* modeling aims to derive θ between feature representations of $\mathcal{X}$ and $\mathcal{Y}$ within a latent space $\mathbb{R}^L$. Specifically, initially specified sequences X^t and Y^t are transformed into $\mathbb{R}^L$, enabling nonlinear computational freedom on their temporal dimensions. Then, a neural network without functional model assumption can derive θ by forming the optimization stream $\mathcal{X} \xrightarrow{\theta} \mathcal{Y}$ in $\mathbb{R}^L$.

The training process uses $\mathcal{X}$ as the input and $\mathcal{Y}$ as the output, indexed through θ, facilitating the concurrent optimization of both dynamical representations. It consequently yields $(\mathcal{X}, \theta, \hat{\mathcal{Y}}_\theta)$ in a sequential association, with each individual representation maintained. The implementation will be introduced in Chapter II.

3.3 Interpretation Complexity

To deal with the often-overlooked dynamical effects, traditional causal inference introduces the concept of "hidden confounder" to enhance model interpretability. For example, the node E in Figure 7 (a) signifies the unobserved individualized characteristics in the scenario depicted in Figure 6.

However, this approach does not necessarily require collecting additional data to identify E. This might lead to an illogical implication: "Our model is biased due to some unknown factors we don't intend to know." Indeed, this strategy employs a solely observational causal variable E to account for the overlooked dynamical aspects of the effect. While E remains unknown, its inclusion can complete the model interpretation. Yet, from the modeling perspective, as illustrated in Figure 7(b), the associative cause $do(A) * E$ remains unknown, failing to provide a modelable relationship for addressing $(\theta, \omega) = \begin{pmatrix} \theta_o \\ \theta_\omega \end{pmatrix}$.


Correlation Model $B_{t+30} = f(do(A_t))$
$do(A)$
$f(do(A))$
the Unobserved Characteristics of Patient $E = \{E_i, E_j, \dots\}$
B
$do(A) * E = \{do(A) * E_i, do(A) * E_j, \dots\}$
$do(A)$
E
Level-II Relation θ_ω
Level-I Relation θ_o
ID
Patient ID $= \{i, j, \dots\}$
Sequences
Sequences
Decode
Encode
$*$
ID
Sequences
ID

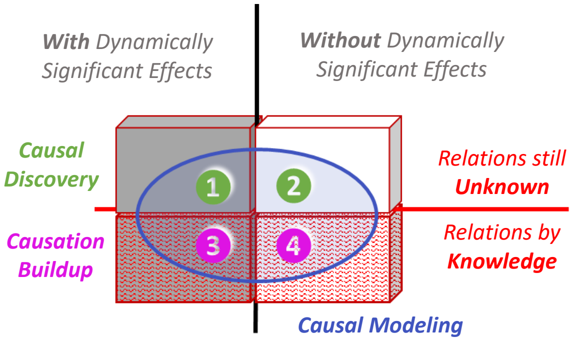
(a) DAG with Hidden Confounder (b) Relation-Indexing Disentanglement (c) Latent Space Architecture of (b)

Figure 7: (a) Traditional causal inference DAG. (b) Hierarchical disentanglement of the dynamical effect through relation-indexing. (c) Autoencoder-based generalized and individualized reconstruction processes.

Incorporating hidden confounders aims to enhance the model's interpretability, though it does not necessarily improve generalizability. In contrast, our *Relation-Oriented* approach bypasses the need to identify cause and effect but simply leverages θ as the index to extract $\hat{\mathcal{Y}}_\theta$, allowing the use of any observed identifier associated with ω (e.g., patient IDs). As demonstrated in section (c), this method effectively disentangles effect representations hierarchically in the latent space, thereby achieving model generalizability.

3.4 Causal Assumptions Reliance

Another consequence of the often-overlooked dynamical effects is the reliance of causal models on foundational assumptions to validate their practical applications. In Figure 8, we categorize causal model applications into four scenarios based on two aspects: 1) depending on whether the predetermined function $f(;\theta)$ is supported by knowledge, they are divided into Causal Discovery and Causation Buildup; 2) the dynamical significance of effects further differentiates them as causality and correlation from the modeling perspective.



	Modeled Relation	**Causal Directionality**
①	Within $\mathbb{R}^O$ Only. Undiscovered Dynamics covered by *Faithfulness Assumption.*	Observational Data Determined. Not Logically Meaningful.
②	Within $\mathbb{R}^O$ Only. Able to align with Knowledge.	Observational Data Determined. Maybe Logically Suggestive.
③	Knowledge Determined. Uncaptured Dynamics covered by *Hidden Confounders* or *Sufficiency Assumption.*	Knowledge Determined.
④	Knowledge Determined.	Knowledge Determined.

Figure 8: Categories of causal learning applications. The left rectangular cube indicates all logically causal relationships, with the potentially modelable ones circled in blue.

Within a generalized-level causation buildup (e.g., the scenario in Figures 6), the dynamical features of the individualized level can be easily overlooked. Based on knowledge, some unobserved entities may be identified as hidden confounders, to enhance model interpretations. Nonetheless, if such identification is not easy, the foundational *Causal Sufficiency* assumption may lead to neglect of these features, presuming that all potential "hidden confounders" have been observed in the system.

On the other hand, causal discovery typically detects relation structures based on observational dependences but excludes the dynamical features of the observables. If these features are not crucial, the captured dependencies can provide valuable insights into the underlying correlations. Otherwise, significant dynamics may be neglected due to the *Causal Faithfulness* assumption, which suggests that the captured observables can fully represent the underlying causal reality.

Furthermore, although the discovered relationships are directional, these directions frequently lack a logical causal implication. Consider X and Y with predetermined directional models $Y = f(X;\theta)$ and $X = g(Y;\phi)$. 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, $\int_X p(x;\theta)dx$ remains constant in this context. Then:

$$\mathcal{I}_{X,Y}(\theta) = \mathbb{E}[(\frac{\partial}{\partial\theta}\log p(X,Y;\theta))^2 \mid \theta] = \int_Y\int_X (\frac{\partial}{\partial\theta}\log p(x,y;\theta))^2 p(x,y;\theta)dxdy$$

$$= \alpha\int_Y (\frac{\partial}{\partial\theta}\log p(y;x,\theta))^2 p(y;x,\theta)dy + \beta = \alpha\mathcal{I}_{Y|X}(\theta) + \beta, \text{with } \alpha,\beta \text{ being constants.}$$

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

The inferred directionality indicates how informatively the observational data distribution can reflect the two predetermined parameters. Consequently, such directionality is unnecessarily logical but could be dominated by the data collection process, with the predominant entity deemed the "cause", consistent with other existing conclusions Reisach (2021); Kaiser (2021). Even when θ and ϕ are predetermined based on knowledge, they might not provide insights for dynamically significant causal relations.

4 Relative Timings in Structural Causality

To visualize the dynamical variations across multiple relative timings, we propose an enhancement to the conventional causal DAGs, which will be utilized in the following sections. Figure 9(a) revisits the example in Figure 7 with hidden-confounder, while the enhancement shown in (b) is carried out through two steps:

1. Consider dynamical effects to integrate necessary relative timings as explicit axes.
2. Use edge lengths to denote timespans for reaching a certain effect magnitude, signified by a static value.

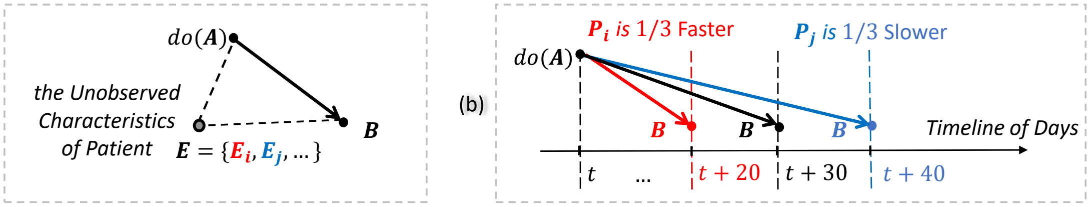


Figure 9: (a) Conventional DAG (Directed Acyclic Graph) with hidden E. (b) Enhanced DAG.

Section 4.1 presents the concept of *inherent bias* via an intuitive example alongside definitive discussions; section 4.2 explores its essential impact on the generalizability of structural models; finally, section 4.3 delves into the advancements and challenges we face in achieving structuralized causal reasoning within AI.

4.1 Scheme of the Inherent Bias

Figure 10(a) shows an example causal structure $\mathcal{B} \leftarrow do(A) \rightarrow \mathcal{C}$ extended from the Figure 9(b) scenario, featuring two medical effects $\mathcal{B}$ and $\mathcal{C}$, on two distinct vital signs B and C, respectively. The primary effect $\mathcal{B}$ is represented as edge $\overrightarrow{AB}$ along $\mathbf{t_1}$, and similar for the side effect $\mathcal{C}$. For simplicity, we assume *nonlinear independence* between $\mathcal{B}$ and $\mathcal{C}$ by fixing the timespan of $\overrightarrow{AC}$ at 10 days for all patients (i.e., $\mathcal{C}$ is dynamically insignificant, denoted as $\mathcal{C} = C$), and focus on predicting a static outcome B, as the average primary effect for the present population. Notably, C can influence B through the *static interaction* edge $\overrightarrow{CB}$.

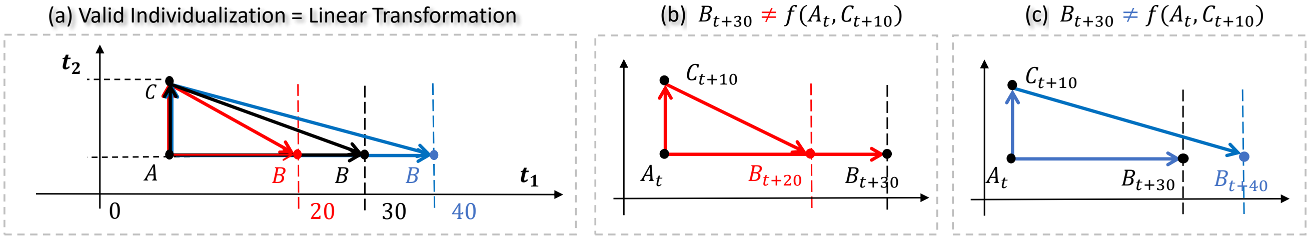


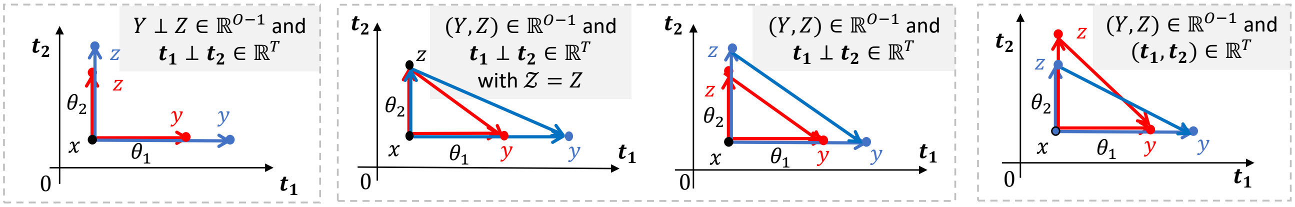
Figure 10: (a) The enhanced DAG with two relative timing axes. (b) and (c): The biased Structural Causal Model (SCM) from the Observed-View, when the timestamp of the static effect B is priorly specified.

From a geometrical view, the triangle over nodes $\{A, B, C\}$ should remain closed for all populations and individuals to represent the same relationship, as supported by the *Causal Markov* condition. Accordingly, the generalization (and also individualization) process can be geometrically viewed as a *linear transformation* of this DAG, depicted as "stretching" the triangle along $\mathbf{t_1}$ at various ratios, as in Figure 10(a).

In conventional SCMs, the status of B is typically identified by setting an average timespan in absolute timing $\mathbf{t}$, for full medicine release along $\overrightarrow{AB}$, say 30 days in this case. As shown in Figure 10(b) and (c), the SCM function fails to shape a valid DAG for individual patients, P_i in red and P_j in blue. Sequential biases would be implied when extending to estimate a sequential outcome like $B^t = B_1, \dots, B_{30}$.

Definition 7. The *Inherent Bias* in SCM assuming *no interactions.*
The *inherent bias* may occur within pre-identified effects if existing: 1) dynamical significance of effects, 2) confounding with interactions across multiple relative timings, and 3) undetectable hierarchy.

Given a structure $\mathcal{Y} \xleftarrow{\theta_1} do(X) \xrightarrow{\theta_2} \mathcal{Z}$, there exist three scenarios regarding interaction between dynamical effects $\mathcal{Y}$ and $\mathcal{Z}$: 1) *no interaction*; 2) only a *static interaction* between them, implying their *linear dependence* and forming a confounding; 3) A *dynamical interaction* between them, implying their *nonlinear dependence* and forming a dynamical confounding. Figure 11 illustrates these definitions.


$Y \perp Z \in \mathbb{R}^{O-1}$ and
$\mathbf{t_1} \perp \mathbf{t_2} \in \mathbb{R}^T$
$(Y, Z) \in \mathbb{R}^{O-1}$ and
$\mathbf{t_1} \perp \mathbf{t_2} \in \mathbb{R}^T$
with $\mathcal{Z} = Z$
$(Y, Z) \in \mathbb{R}^{O-1}$ and
$\mathbf{t_1} \perp \mathbf{t_2} \in \mathbb{R}^T$
$(Y, Z) \in \mathbb{R}^{O-1}$ and
$(\mathbf{t_1}, \mathbf{t_2}) \in \mathbb{R}^T$
$\mathbf{t_1}$
$\mathbf{t_2}$
θ_1
θ_2
x
y
z
0

(1) *No Interaction* between $\mathcal{Y}$ and $\mathcal{Z}$, defined as $\theta_1 \perp \theta_2 \in \mathbb{R}^H$

(2) *Linear Dependence* between $\mathcal{Y}$ and $\mathcal{Z}$ with *Static Interaction*, defined as $\theta_1 \propto \theta_2 \in \mathbb{R}^H$

(3) *Nonlinear Dependence* with *Dynamical Interaction*, defined as $(\theta_1, \theta_2) \in \mathbb{R}^H$

Figure 11: Illustrative definitions of *Interaction, Dependence, and Confounding* between dynamical effects $\mathcal{Y} = \langle Y, \mathbf{t_1} \rangle$ and $\mathcal{Z} = \langle Z, \mathbf{t_2} \rangle$ within $\mathbb{R}^{O-1} \cup \mathbb{R}^T$, where $\{x, y, z\}$ indicate certain static values of $\{X, Y, Z\}$.

Remark 7. The ***inherent bias*** remains to exist within AI-based SCMs, which assume *nonlinear independences* (i.e., scenarios 1 and 2) among the captured dynamical effects.

The second scenario, defined as $\theta_1 \propto \theta_2 \in \mathbb{R}^H$, serves as a special case of the third, $(\theta_1, \theta_2) \in \mathbb{R}^H$. This presupposes the dynamically significant effects do not exhibit nonlinear temporal dependence, among their temporal dimensions. Indeed, it is practically challenging to distinguish the three scenarios from data alone without prior knowledge. Thus, making the assumption $(\theta_1, \theta_2) \in \mathbb{R}^H$ becomes a cautious default setting.

To construct an SCM, the effects $\mathcal{Y}$ and $\mathcal{Z}$ are initially identified as Y^t and Z^t in absolute timing $\mathbf{t}$. If satisfying $\theta_1 \propto \theta_2 \in \mathbb{R}^H$, AI models like inverse RNNs can accurately capture $\vartheta = \overrightarrow{\theta_1\theta_2}$ by constructing $do(X) = f\big((Y, Z)^t; \vartheta\big)$ using their associative identification $(Y, Z)^t = \big((Y, Z)_1, \ldots, (Y, Z)_t\big)$. However, under a more general condition $(\theta_1, \theta_2) \in \mathbb{R}^H$, it may introduce ***inherent bias***, as neglecting dynamical interactions among dynamical effects, as highlighted in Remark 7. Its scheme is similar to Definition 7, where the conventional SCM assumes linearly independent static effects, i.e., $\mathcal{Y} = Y$, $\mathcal{Z} = Z$, and $\theta_1 \perp \theta_2$.

It is essential to adopt a two-step *relation-indexed learning* to sequentially obtain models $\mathcal{Y} = f_1(do(X); \theta_1)$ and $\mathcal{Z} = f_2(do(X) \mid \mathcal{Y}; \theta_2)$. Without this approach, when modeling large-scale causal structures, the inherent biases can accumulate within AI models, compromising robustness and leading to irrational outputs.

4.2 Inherently Restricted Generalizability

To address the inherent biases within conventional SCMs, traditional causal inference uses various methods to perform "de-confounding", such as propensity score matching Benedetto (2018), backdoor adjustment Pearl (2009), etc. They aim to cut off the static interactions between dynamical effects, although without explicitly recognizing these dynamics. These techniques typically rely on intended tailoring for each specific application. Given the black-box nature and the large scale of AI models, such manual identification and adjustment approaches become increasingly impractical for modern causal learning inquiries.

Moreover, they primarily deal with linear dependence only, to adapt to statistical linear models, which may not contribute to dynamical generalizability. Subsequently, we employ a practical example to illustrate how effect identifications inherently hinder SCM's generalizability, under the general condition $(\theta_i, \theta_j) \in \mathbb{R}^H$.

Figure 12 displays a 3D view enhanced DAG, where Δt and $\Delta\tau$ signify actual time spans for the present population, to support their causal reasoning represented by this structure. For the triangle $SA'B'$, as each unit of effect from S delivered to A' (spent $\Delta\tau$), it immediately starts to impact B' through $\overrightarrow{A'B'}$ (Δt needed); meanwhile, the next unit begins generation at S. This dual action runs concurrently until S's effect fully reaches B', represented as the single edge $\overrightarrow{SB'}$ within this SCM.

Due to the equation $\overrightarrow{SB'} = \overrightarrow{SA'} + \overrightarrow{A'B'}$, specifying the time span of $\overrightarrow{SB'}$ inherently determines the $\Delta t : \Delta\tau$ ratio based on the current population's performance, thereby fixing the shape of the ASB' triangle in the DAG space. If we focus solely on the accuracy of the estimated mean effect for this population, the SCM function $B' = f(A, C, S)$ may be effective. However, given that the preset $\Delta t : \Delta\tau$ ratio is not universally applicable, the generalizability of the established SCM to other populations becomes questionable.

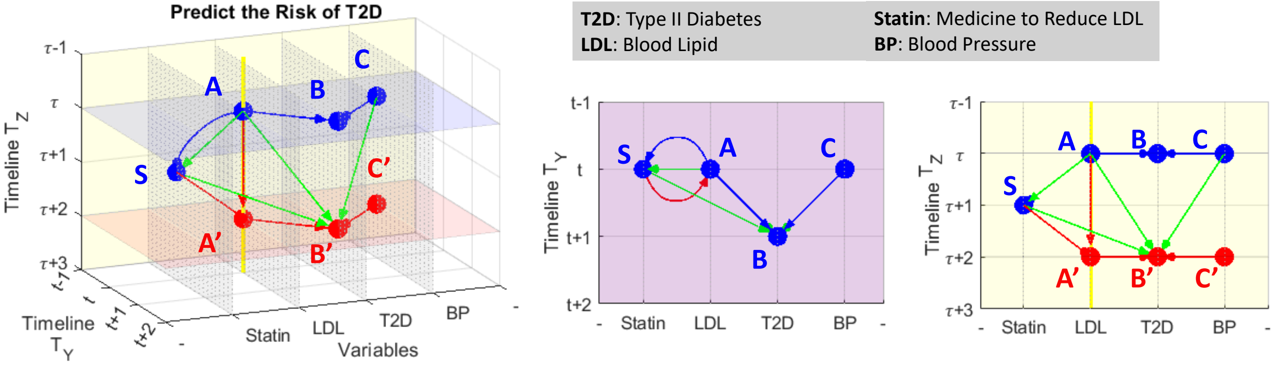
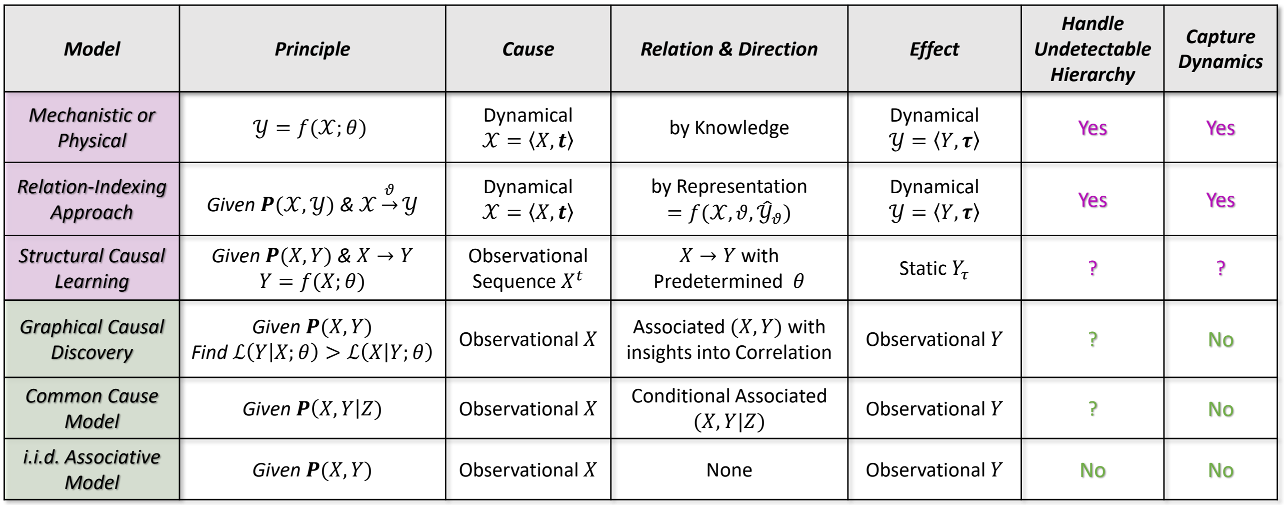


Figure 12: DAG space with two relative timing axes T_Y and T_Z. The SCM $B' = f(A, C, S)$ is to evaluate the effect of using S to reduce T2D risks at B'. On T_Y, the step Δt from t to $(t+1)$ allows A and C to fully influence B; the step $\Delta\tau$ on T_Z from $(\tau+1)$ to $(\tau+2)$ let S fully release to forward status A to A'.

4.3 Developments Toward Causal Reasoning AI

To pursue causal reasoning in machine learning, our modeling techniques have evolved from capturing mere associations to observational correlations, ultimately advancing to build structural causal models spinning the counterfactual temporal space $\mathbb{R}^T$. Figure 13 summarizes this evolution in an upward trajectory.



Model	*Principle*	*Cause*	*Relation & Direction*	*Effect*	*Handle Undetectable Hierarchy*	*Capture Dynamics*
Mechanistic or Physical	$\mathcal{Y} = f(\mathcal{X}; \theta)$	Dynamical $\mathcal{X} = \langle X, \boldsymbol{t} \rangle$	by Knowledge	Dynamical $\mathcal{Y} = \langle Y, \boldsymbol{\tau} \rangle$	Yes	Yes
Relation-Indexing Approach	Given $\mathbf{P}(\mathcal{X}, \mathcal{Y})$ & $\mathcal{X} \overset{\vartheta}{\rightarrow} \mathcal{Y}$	Dynamical $\mathcal{X} = \langle X, \boldsymbol{t} \rangle$	by Representation $= f(\mathcal{X}, \vartheta, \hat{\mathcal{Y}}_\vartheta)$	Dynamical $\mathcal{Y} = \langle Y, \boldsymbol{\tau} \rangle$	Yes	Yes
Structural Causal Learning	Given $\mathbf{P}(X, Y)$ & $X \rightarrow Y$ $Y = f(X; \theta)$	Observational Sequence X^t	$X \rightarrow Y$ with Predetermined θ	Static Y_τ	?	?
Graphical Causal Discovery	Given $\mathbf{P}(X, Y)$ Find $\mathcal{L}(Y\|X; \theta) > \mathcal{L}(X\|Y; \theta)$	Observational X	Associated (X, Y) with insights into Correlation	Observational Y	?	No
Common Cause Model	Given $\mathbf{P}(X, Y\|Z)$	Observational X	Conditional Associated $(X, Y\|Z)$	Observational Y	?	No
i.i.d. Associative Model	Given $\mathbf{P}(X, Y)$	Observational X	None	Observational Y	No	No

Figure 13: Simple taxonomy of models (partially refer to Scholkopf (2021) Table 1), from more data-driven upward to more knowledge-driven. "?" means depending on the practice.

Given AI's capability to learn nonlinear dynamics, the present challenge is incorporating the underlying dynamical interactions within the causal knowledge structure. Considering the risk of introducing inherent biases, finding a new modeling paradigm is crucial to realizing causal knowledge-aligned AI. Physical models, explicitly incorporated in temporal dimensional computation, may offer valuable insights into this prospect.

Under the observational i.i.d. assumption, initial models only approximate associations, proved unreliable for causal reasoning Pearl et al. (2000); Peters et al. (2017). Subsequently, the common cause principle highlights the significance of the nontrivial condition, to distinguish a relationship from statistical dependencies Dawid

(1979); Geiger (1993), providing a basis for constructing graphical models Peters et al. (2014). The initial graphical model relies on conditional dependencies to construct Bayesian networks, with limited causal relevance Scheines (1997). Then, causally significance emphasizes the capability of addressing counterfactual queries Scholkopf (2021), like the structural equation models (SEMs) and functional causal models (FCMs) Glymour et al. (2019); Elwert (2013), which leverage prior knowledge to establish causal structures.

State-of-the-art deep learning on causality encodes the discrete, DAG-structural constraint into continuous optimization functions Zheng et al. (2018; 2020); Lachapelle et al. (2019), enabling advanced efficiency, but without noticeable generalizability, evident from the restricted successes in applications like the neural architecture search (NAS) Luo (2020); Ma (2018). This is reasonable since the neglected interactions among relative timings can lead to inherent biases amplified through complex structures to become significant. Scholkopf (2021) summarized our confronting key challenges toward generalizable causal-reasoning AI: 1) limited model robustness, 2) insufficient model reusability, and 3) inability to handle data heterogeneity (i.e., undetectable hierarchies). They are intrinsically linked to the demonstrated inherent biases.

Chapter II: Realization of Proposed Relation-Oriented Paradigm

This chapter introduces the proposed *Relation-Indexed Representation Learning* (RIRL) method, a baseline realization of the raised *Relation-Oriented* modeling paradigm. RIRL primarily focuses on autonomously identifying dynamical effects, in the form of relation-indexed representations in the latent space. In the context of structural modeling, RIRL enables hierarchical disentanglement of effects, according to given DAGs, as a manner of realizing dynamical generalizability across undetectable levels within knowledge. As a baseline realization, RIRL is suitable for applications with mature structural causal knowledge, and plenty of data to support neural network training on each known causal relationship.

First, Section 5 details the technique for extracting relation-indexed representations. Then, building on this, Section 6 presents the RIRL method of establishing structural causal models in the latent space. Lastly, Section 7 provides experiments to validate RIRL's efficacy in autonomously identifying effects.

5 Relation-Indexed Representation

In the relationship $\mathcal{X} \to \mathcal{Y}$, we define dynamical $\mathcal{X} = \langle X, t \rangle \in \mathbb{R}^{d+1} \subseteq \mathbb{R}^O$ and $\mathcal{Y} = \langle Y, \tau \rangle \in \mathbb{R}^{b+1} \subseteq \mathbb{R}^O$, given their solely observational variables, $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^b$. $\mathcal{X}$ is observed as a data sequence, represented by $X^t = X_1, \ldots, X_t$ with a pre-determined length l_x. For clarity, hereafter in this chapter, its instance x^t will be considered as a $(d * l_x)$-dimensional vector, denoted by $\overrightarrow{x}$ (or x for briefty). Similarly, $\mathcal{Y}$ is observed as the data sequence Y^t with a pre-determined length l_y, and its instance is referred to as a $(b * l_y)$-dimensional vector $\overrightarrow{y}$ (or y for briefty).

The relation-indexed representation aims to formulate $(\mathcal{X}, \theta, \hat{\mathcal{Y}}_\theta)$ in the latent space $\mathbb{R}^L$, beginning with an *initialization* to transform X^t and Y^t to be latent space features. For the sake of 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 modeling process is to optimize the neural network function $f(;\theta)$ in $\mathbb{R}^L$, with $\mathcal{H}$ as its input and $\mathcal{V}$ as the output. This process simultaneously refines $\mathcal{H}$, θ, and $\mathcal{V}$, for ultimately achieving $(\mathcal{H}, \theta, \hat{\mathcal{V}}_\theta) = (\mathcal{X}, \theta, \hat{\mathcal{Y}}_\theta)$. The refining will present as the distance minimization between $\mathcal{H}$ and $\mathcal{V}$ within $\mathbb{R}^L$. Consequently, the dimensionality L of the latent feature space must satisfy $L \geq rank(\mathcal{X}, \theta, \mathcal{Y})$, raising a technical challenge that L is larger than the dimensionality of $\overrightarrow{x}$ or $\overrightarrow{y}$.

Remark 8. The variable *initialization* necessitates a *higher-dimensional* representation autoencoder.

5.1 Higher-Dimensional Autoencoder

Autoencoders are commonly used for dimensionality reduction, especially in structural modeling that involves multiple variables Wang (2016). In contrast, RIRL aims to model individual causal relationships sequentially within a higher-dimensional latent space $\mathbb{R}^L$, as to hierarchically construct the entire causal structure. As illustrated in Figure 14, the designed autoencoder architecture is featured by the symmetrical *Expander* and *Reducer* layers (source code is available [1]). The Expander magnifies the input vector $\overrightarrow{x}$ by capturing its higher-order associative features, while the Reducer symmetrically diminishes dimensionality and reverts to its initial state. For precise reconstruction, the *invertibility* of these processes is essential.

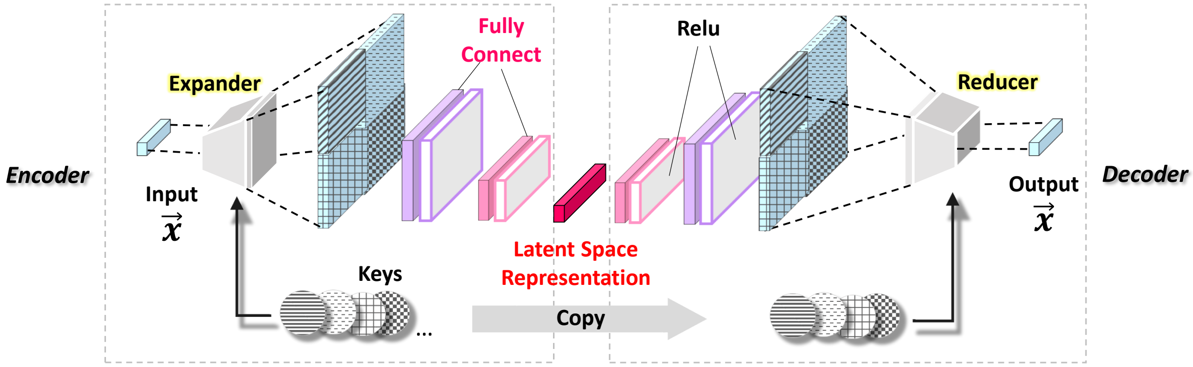


Figure 14: *Invertible* autoencoder architecture for extracting *higher-dimensional* representations.

The Expander showcased in Figure 14 implements a *double-wise* expansion. Here, every duo of digits from $\overrightarrow{x}$ is encoded into a new digit using an association with a random constant, termed the *Key*. This *Key* is generated by the encoder and replicated by the decoder. Such pairwise processing of $\overrightarrow{x}$ expands its length from $(d * l_x)$ to be $(d * l_x - 1)^2$. By leveraging multiple *Keys* and concatenating their resultant vectors, $\overrightarrow{x}$ can be considerably expanded, ready for the subsequent dimensionality-reduced representation extraction.

The four blue squares with unique grid patterns represent expansions by four distinct *Keys*, with the grid patterns acting as their "signatures". Each square symbolizes a $(d * l_x - 1)^2$ length vector. Similarly, higher-order expansions, like *triple-wise* across three digits, can be achieved with adapted *Keys*.

Figure 15 illustrates the encoding and decoding processes within the Expander and Reducer, targeting the digit pair (x_i, x_j) for $i \neq j \in 1, \ldots, d$. The Expander function is defined as $\eta_\phi(x_i, x_j) = x_j \otimes exp(s(x_i)) + t(x_i)$, which hinges on two elementary functions, $s(\cdot)$ and $t(\cdot)$. The *Key* parameter, ϕ, embodies their weights, $\phi = (w_s, w_t)$. Specifically, the Expander morphs x_j into a new digit y_j utilizing x_i as a chosen attribute. In contrast, the Reducer symmetrically uses the inverse function η_ϕ^{-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.

---

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

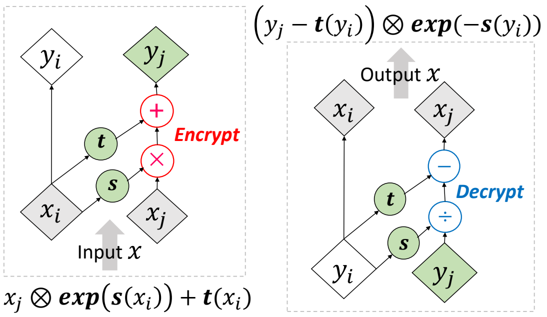

y_i
y_j
t
s
Encrypt
x_i
x_j
Input x
$x_j \otimes \boldsymbol{exp}(\boldsymbol{s}(x_i)) + \boldsymbol{t}(x_i)$
$(y_j - \boldsymbol{t}(y_i)) \otimes \boldsymbol{exp}(-\boldsymbol{s}(y_i))$
Output x
x_i
x_j
Decrypt
t
s
y_i
y_j

Figure 15: Expander (left) and Reducer (right).

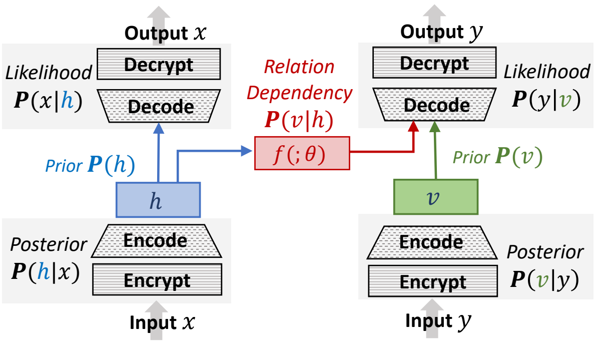


Figure 16: Relationship model architecture.

5.2 Optimization Steps

Consider instances x and y of $\mathcal{X}$ and $\mathcal{Y}$, with corresponding representations h and v in $\mathbb{R}^L$. The latent dependency $\mathbf{P}(v|h)$ is used to train the relation function $f(;\theta)$, as illustrated in Figure 16. In each iteration, the modeling process undergoes three optimization steps:

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.

During this process, the values of h and v are iteratively adjusted to reduce their distance in $\mathbb{R}^L$, with $f(;\theta)$ serving as a bridge to span the distance. Here, the hyper-dimensional variable $\theta \in \mathbb{R}^H$ acts as the index, guiding the output of $f(;\theta)$ to fulfill associated representations $(\mathcal{H}, \theta, \hat{\mathcal{V}}_\theta)$. From $\hat{\mathcal{V}}_\theta$, the effect component $\hat{\mathcal{Y}}_\theta$, also the causal representation, can be reconstructed. Within this system, for each effect, a series of such relation functions $\{f(;\theta)\}$ is maintained, indexing diverse levels of causal inputs for sequentially building the structural model.

6 RIRL: Building Structural Models in Latent Space

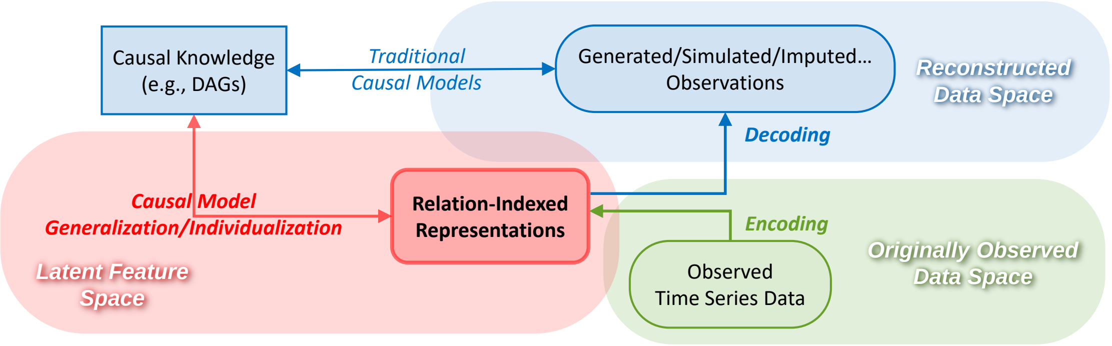

Causal Knowledge (e.g., DAGs)
Traditional Causal Models
Generated/Simulated/Imputed… Observations
Reconstructed Data Space
Decoding
Causal Model Generalization/Individualization
Relation-Indexed Representations
Encoding
Originally Observed Data Space
Latent Feature Space
Observed Time Series Data

Figure 17: How Relation-Indexed Representation Learning (RIRL) contributes to traditional models.

By sequentially constructing relation-indexed representations for each pairwise relationship within the causal DAG, we can achieve the hierarchically disentangled representation for each node, according to its levels defined by the global structure. Simultaneously, the entire structualized causality has also been constructed. Subsequently, section 6.1 details the method for stacking relation-indexed representations, enabling the construction of higher-level representations based on previously established lower-level ones; section 6.2 provides the complete factorization process for hierarchical disentanglement; finally, section 6.3 discusses a causal discovery algorithm within the latent space among initialized variable representations.

Figure 17 demonstrates how the RIRL method can encapsulate the black-box nature of AI within the latent space while simultaneously generating interpretable observations. This characteristic can be utilized to enhance conventional *Observation-Oriented* models, for instance, by simulating counterfactual values on demands. Meanwhile, in the latent space, these cryptic representations, although opaque to human interpretation, play a crucial role in achieving model generalization and individualization. These processes are latently managed by AI and remain exclusive to human comprehension.

6.1 Stacking Hierarchical Representations

A structural relationship can be represented by a causal graph, denoted as G. To construct models in the latent space, the latent dimensionality L must be sufficiently large to adequately represent G. Let's denote a data matrix augmented by all observational attributes in G as $\mathbf{X}$. Given the need to include informative relations $\{\theta\}$ for the edges in G, it is essential that $L > rank(\mathbf{X}) + T$, where T indicates the number of dynamically significant variables (i.e., nodes) within G.

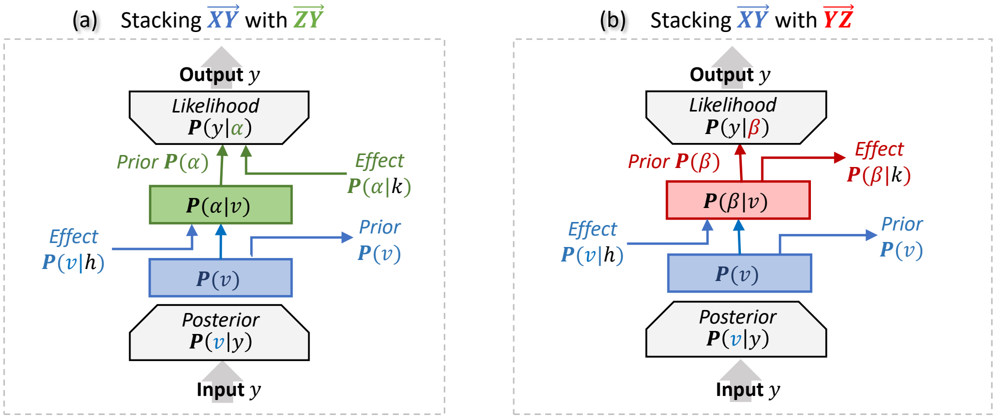


Figure 18: Stacking relation-indexed representations to construct hierarchy.

The PCA principle posits that the space $\mathbb{R}^L$ learned by the autoencoder is spanned by the top principal components of $\mathbf{X}$ Baldi (1989); Plaut (2018); Wang (2016). Hypothetically, reducing L below $rank(\mathbf{X})$ may yield a less adequate but causally more significant latent space through better alignment of dimensions Jain (2021) (Further exploration in this direction is warranted). Bypassing a deep dive into dimensionality boundaries, we rely on empirical fine-tuning for the experiments in this study (reducing L from 64 to 16).

Consider a causal structural among $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$, with their corresponding representations $\{\mathcal{H}, \mathcal{V}, \mathcal{K}\} \in \mathbb{R}^L$ initialized by three autoencoders, respectively. Figure 18 illustrates the hierarchical representations buildup. Here, two stacking scenarios are displayed based on varying causal directions. With the established $\mathcal{X} \rightarrow \mathcal{Y}$ relationship in $\mathbb{R}^L$, the left-side architecture finalizes the $\mathcal{X} \rightarrow \mathcal{Y} \leftarrow \mathcal{Z}$ structure, while the right-side focuses on $\mathcal{X} \rightarrow \mathcal{Y} \rightarrow \mathcal{Z}$. Through the addition of a representation layer, hierarchical disentanglement is formed, allowing for various input-output combinations (denoted as $\mapsto$) according to specific requirements.

For example, on the left, $\mathbf{P}(v|h) \mapsto \mathbf{P}(\alpha)$ represents the $\mathcal{X} \rightarrow \mathcal{Y}$ relationship, whereas $\mathbf{P}(\alpha|k)$ implies $\mathcal{Z} \rightarrow \mathcal{Y}$. Conversely, on the right, $\mathbf{P}(v) \mapsto P(\beta|k)$ denotes the $\mathcal{Y} \rightarrow \mathcal{Z}$ relationship with $\mathcal{Y}$ as input. Meanwhile, $\mathbf{P}(v|h) \mapsto P(\beta|k)$ captures the causal sequence $\mathcal{X} \rightarrow \mathcal{Y} \rightarrow \mathcal{Z}$.

6.2 Factorizing the Effect Disentanglement

Consider $\mathcal{Y} = \langle Y, \tau \rangle \in \mathbb{R}^{b+1} \subseteq \mathbb{R}^O$ having a n-level hierarchy, with each level built up using a representation function, labeled as g_i for the i-th level. For simplicity, here, we use ω_i to represent the i-th level component of $\mathcal{Y}$ in the latent space $\mathbb{R}^L$, while its counterpart in $\mathbb{R}^{b+1}$ is denoted as Ω_i (i.e., $\hat{\mathcal{Y}}$ at the i-th level). Let the feature vector ω_i in $\mathbb{R}^L$ primarily spans a sub-dimensional space, $\mathbb{R}^{L_i}$, resulting in the spatial disentanglement sequence $\{\mathbb{R}^{L_1}, \ldots, \mathbb{R}^{L_i}, \ldots, \mathbb{R}^{L_n}\}$, which hierarchically represents $\mathcal{Y}$ with n components. Function g_i maps from $\mathbb{R}^{b+1}$ to $\mathbb{R}^{L_i}$, taking into account features from all previous levels as attributes. This gives us:

$$\mathcal{Y} = \sum_{i=1}^{n} \Omega_i, \;\text{ where } \Omega_i = g_i\big(\omega_i;\; \Omega_1, \ldots, \Omega_{i-1}\big) \text{ with } \Omega_i \in \mathbb{R}^{b+1} \text{ and } \omega_i \in \mathbb{R}^{L_i} \subseteq \mathbb{R}^L \tag{1}$$

In the context of a purely observational hierarchy, with $\mathcal{Y}$ substituted by $Y \in \mathbb{R}^b$, The example depicted in Figure 2 (b) can be interpreted as follows: Consider three feature levels represented as $\omega_1 \in \mathbb{R}^{L_1}$, $\omega_2 \in \mathbb{R}^{L_2}$, and $\omega_3 \in \mathbb{R}^{L_3}$. For simplicity, assume the subspaces are mutually exclusive, such that $L = L_1 + L_2 + L_3$. In the latent space, the triplet $(\omega_1, \omega_2, \omega_3) \in \mathbb{R}^L$ comprehensively depicts the image. Their observable counterparts, Ω_1, Ω_2, and Ω_3, are three distinct full-scale images, each showcasing different content. For example, Ω_1 emphasizes finger details, while the combination $\Omega_1 + \Omega_2$ reveals the entire hand.

6.3 Causal Discovery in Latent Space

Algorithm 1 outlines the heuristic procedure for investigating edges among the initialized variable representations. We use Kullback-Leibler Divergence (KLD) as a metric to evaluate the strength of causal relationships. Specifically, as depicted in Figure 16, KLD evaluates the similarity between the relation output $\mathbf{P}(v|h)$ and the prior $\mathbf{P}(v)$. Lower KLD values indicate stronger causal relationships due to closer alignment with the ground truth. Conversely, while Mean Squared Error (MSE) is a frequently used evaluation metric, its sensitivity to data variances Reisach (2021) leads us to utilize it as a supplementary measure in this study.

Algorithm 1: Latent Space Causal Discovery

```
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 $\beta \to n$
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

Figure 19 illustrates the causal structure discovery process in latent space over four steps. Two edges, (e_1 and e_3), are sequentially selected, with e_1 setting node B as the starting point for e_3. In step 3, edge e_2 from A to C is deselected and reassessed due to the new edge e_3 altering C's present causal conditions. The final DAG represents the resulting causal structure.

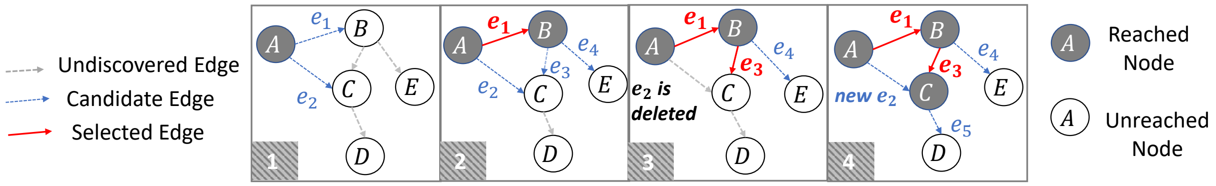


Figure 19: An example of causal discovery in the latent space.

7 Efficacy Validation Experiments

The experiments aim to validate the efficacy of the RIRL method from three aspects: 1) the performance of the proposed higher-dimensional representations, evaluated by reconstruction accuracy, 2) the construction of a clear effect hierarchy through the stacking of relation-indexed representations, and 3) the identification of DAG structures within the latent space through discovery. A full demonstration of the conducted experiments in this chapter is available online [2], while with two primary limitations detailed as follows:

Firstly, the dataset employed in this study may not be the most suitable for evaluating the effectiveness of RIRL. Ideally, real-world data featuring rich structuralized causality across multiple relative timings, like clinical records, would be preferable. However, due to practical constraints, access to such optimal data is limited for this study, leading us to use the current synthetic data and focus solely on feasibility verification. For experimental validation regarding the inherent bias, please refer to prior research Li et al. (2020).

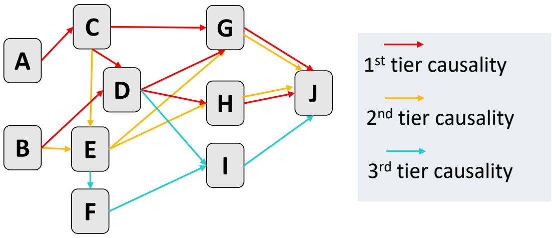
Secondly, the time windows designated for cause and effect, l_x and l_y, are fixed at 10 and 1, respectively. This constraint arose from an initial oversight in the experimental design stage, wherein the pivotal role of effect dynamics has not been fully recognized, consequently limited by the RNN pattern. It manifests as

---

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

restricted successes in building causal chains like $\mathcal{X} \rightarrow \mathcal{Y} \rightarrow \mathcal{Z}$; while the model can adeptly capture single-hop causality, it struggles with multi-hop ones since the dynamics in $\mathcal{Y}$ have been segmented by $l_y = 1$. However, extending the length of l_y does not pose a significant technical challenge to future works.

7.1 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 20: Hydrological causal DAG: routine tiers organized by descending causal strength.

The dataset chosen for our experiments is 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 (2020); Kratzert (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 20 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 (1st tier) significantly impacts rapid streamflow peaks, followed by lateral flow (2nd tier). Baseflow dynamics (3rd tier) have a subtler influence. Our causal discovery experiments aim to reveal these underlying tiers.

7.2 Higher-Dimensional Variable Representation Test

In this test, we have a total of ten variables (i.e., nodes), with each requiring an individual autoencoder for initialization. Table 1 lists the statistical characteristics of their post-scaled (i.e., normalized) attributes, along with their autoencoders' reconstruction accuracies. Accuracy is assessed in the root mean square error (RMSE), where a lower RMSE indicates higher accuracy for both scaled and unscaled data.

The task is challenging due to the limited dimensionalities of the ten variables - 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.

7.3 Hierarchical Disentanglement Test

Table 2 provides the performance comparison of stacking relation-indexed representations on each node. The term "single-effect" is to describe the accuracy of a specific effect node when reconstructed from a single cause

Table 1: Characteristics of node attributes and their variable representation test results.

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

node (e.g., $B \to D$ and $C \to D$), and "full-effect" for the accuracy when all its cause nodes are stacked (e.g., $BC \to D$). To provide context, we also include baseline performance scores based on the initialized variable representations. During the relation learning process, the effect node serves two purposes: it maintains its own accurate representation (as per optimization no.2 in 5.2) and helps reconstruct the relationship (as per optimization no.1 in 5.2). Both aspects are evaluated in Table 2.

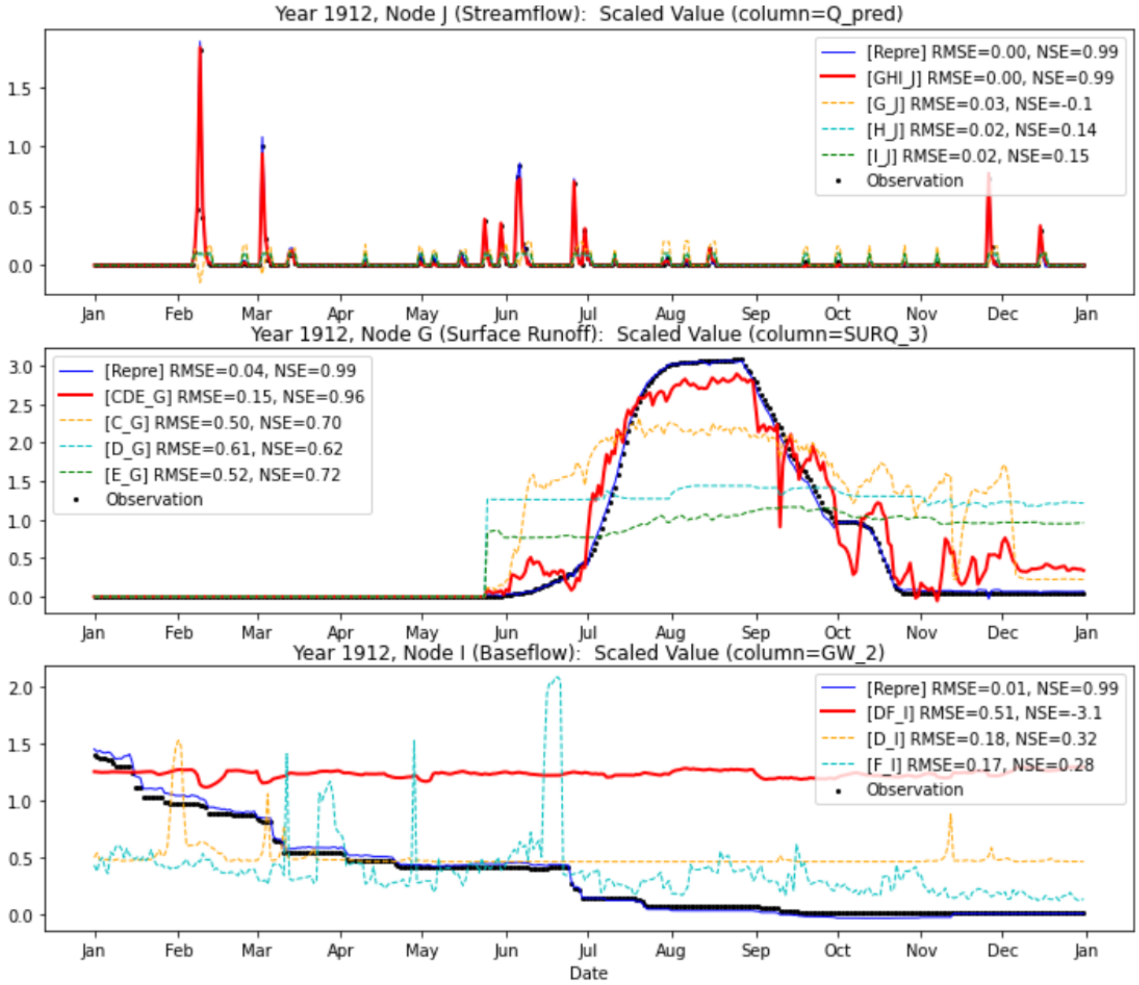


Figure 21: Reconstructed dynamical effects, via hierarchically stacked relation-indexed representations.

The KLD metrics in Table 2 indicate the strength of learned causality, with a lower value signifying stronger. For instance, 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-effect" are moderate compared to their "single-effect"

Table 2: Effect Reconstruction Performances of RIRL sorted by effect nodes.

Result Node	Variable Representation (Initial)			Cause Node	Variable Representation (in Relation Learning)			Relationship Reconstruction			
	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

Table 3: Brief summary of the latent space causal discovery test.

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

scores, suggesting a lack of informative associations among the cause nodes. In contrast, for nodes G and H, lower "full-effect" 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 $C \rightarrow G$ strength to $CDE \rightarrow G$ suggests that C is the primary contributor to this causal relationship.

Figure 21 showcases reconstructed time series, for the effect nodes J, G, and I, in the same synthetic year to provide a straightforward overview of the hierarchical representation performances. Here, black dots represent the ground truth; the blue line indicates reconstruction via the initial variable representation, and the "full-effect" 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 initial variable representation closely aligns with the ground truth, as shown in Figure 21, attesting to the efficacy of our proposed autoencoder architecture. As expected, the "full-effect" performs better than the "single-effect" 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.

7.4 Latent Space Causal Discovery Test

The discovery test initiates with source nodes A and B and proceeds to identify potential edges, 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 order in which existing edges are discovered, along with the corresponding KLD sums and gains after each edge is included. Color-coding in the cells corresponds to Figure 20, indicating tiers of causal routines. The arrangement underscores the efficacy of this latent space discovery approach.

A comprehensive list of candidate edges evaluated in each discovery round is provided in Table 4 in Appendix A. For comparative purposes, we also performed a 10-fold cross-validation using the conventional FGES discovery method; those results are available in Table 5 in Appendix A.

8 Conclusions

This paper introduces a dimensionality framework from a *Relation-Oriented* perspective to decompose our cognitive space, where relational causal knowledge is stored. Specifically, it conceptualizes the unobservable relations between cause and effect as informative variables in $\mathbb{R}^H$; and the causal structure of dynamics in knowledge, represented by the enhanced DAG, is accommodated by the counterfactual space $\mathbb{R}^T$, across multiple relative timing axes with nonlinear dependence. It highlights the key oversights of the current *Observation-Oriented* paradigm, which relies on the observational i.i.d. assumption and is confined to $\mathbb{R}^O$.

The traditional causal inference, adopting a *Relation-Oriented* viewpoint, identifies the underlying causal structures across relative timings but overlooks the $\mathbb{R}^T$ space due to neglecting temporal nonlinearities, i.e., the dynamics. Under the *Observation-Oriented* paradigm, contemporary causal learning is often challenged by incompletely captured dynamical effects without considering the indexing role of unobservable relations lying in $\mathbb{R}^H$. In the case of LLMs, while AI techniques enable the autonomous identification of dynamical effects, they often neglect their interactions, which are emphasized as causal structures in causal inference.

Recalling the queries presented in the Introduction, we systematically summarize these application-related restrictions in our pursuit of AGI, and offer new insights:

- *Firstly*, challenges for causal inference models primarily arise from overlooking dynamics, due to their linear modeling constraints. This oversight leads to various compensatory efforts, such as introducing hidden confounders and relying on the causal sufficiency assumption. Causal DAGs inherently provide a *Relation-Oriented* view; with the proposed enhancement incorporating them into the counterfactual $\mathbb{R}^T$ space, they can provide essential support for illustrating structuralized dynamics.
- *Secondly*, our knowledge inherently contains hierarchical levels due to hidden relations $\omega \in \mathbb{R}^H$, which necessitates generalizability of models. For AI-based causal models, the main challenge lies in incorporating the underlying structure of dynamics to achieve dynamical generalizability. The new paradigm we propose introduces a relation-indexing methodology, enabling the autonomous construction of causal structures by sequentially extracting causal representations.
- *Thirdly*, while existing language models have made strides in generalizability through meta-learning, they are still limited to absolute timing within $\mathbb{R}^O$, implicitly assuming nonlinear independence among temporal dimensions. Additionally, their neglect of extracting informative θ prevents them from truly "understanding" the captured relationships. However, LLMs have demonstrated the effectiveness of meta-learning in addressing temporal dimensional hierarchies, suggesting a promising prospect for *Relation-Oriented* meta-learning in advancing towards AGI.

We also introduce a baseline implementation of the *Relation-Oriented* paradigm, primarily to validate the efficacy of the "relation-indexing" methodology in implementing causal representations and constructing knowledge-aligned causal structures. Similar thoughts have been attempted in domains with well-established causal knowledge, such as the hierarchical temporal memory in neuroscience Wu (2018). The pursuit of AGI is a historically extensive and complex endeavor, requiring a wide array of knowledge-aligned AI model constructions. This study aims to provide foundational insights for future developments in this field.

References

Daniel L Alkon, Howard Rasmussen. A spatial-temporal model of cell activation. *Science*, 239(4843):998–1005, 1988.

Natalia Andrienko, et al. 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.

Pierre Baldi, Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Umberto Benedetto, et al. Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, 53(6):1112–1117, 2018.

Seana Coulson, et al. Understanding timelines: Conceptual metaphor and conceptual integration. *Cognitive Semiotics*, 5(1-2):198–219, 2009.

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.

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

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.

Dan Geiger, et al. Logical and algorithmic properties of conditional independence and graphical models. *The annals of statistics*, 21(4):2001–2021, 1993.

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, et al. 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, et al. Modelling non-linear economic relationships. *OUP Catalogue*, 1993.

Sander Greenland, et al. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

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

Aapo Hyvärinen, et al. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Saachi Jain, et al. A mechanism for producing aligned latent spaces with autoencoders. *arXiv preprint arXiv:2106.15456*, 2021.

Marcus Kaiser, et al. Unsuitability of notears for causal graph discovery. *arXiv:2104.05441*, 2021.

Frederik Kratzert, et al. 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, et al. 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, et al. 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, et al. 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.

Mariusz Maziarz. A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues*, 8(2):86–105, 2015.

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

Mohammed Ombadi, et al. 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.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

PGMadhavan. Static dynamical machine learning – what is the difference? `https://www.datasciencecentral.com/static-dynamical-machine-learning-what-is-the-difference/`, 2016.

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.

Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv:1804.10253*, 2018.

Alexander G Reisach, et al. Beware of the simulated dag! varsortability in additive noise models. *arXiv preprint arXiv:2102.13647*, 2021.

Schaeffer Rylan, et al. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

Pedro Sanchez, et al. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

Richard Scheines. An introduction to causal inference. 1997.

Bernhard Scholkopf, et al. Toward causal representation learning. *IEEE*, 109(5):612–634, 2021.

Charles H Shea, et al. 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.

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

Monica G Turner. Spatial and temporal analysis of landscape patterns. *Landscape ecology*, 4:21–30, 1990.

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

Yasi Wang, et al. Auto-encoder based dimensionality reduction. 184:232–242, 2016.

Naftali Weinberger and Colin Allen. Static-dynamic hybridity in dynamical models of cognition. *Philosophy of Science*, 89(2):283–301, 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, et al. Hierarchical temporal memory method for time-series-based anomaly detection. *Neurocomputing*, 273:535–546, 2018.

Gabriele Wulf, et al. 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 of Causal Discovery

Table 4: The Complete Results of Heuristic Causal Discovery in 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