# Temporal Knowledge Graph Extrapolation via Causal Subhistory Identification

**Kai Chen**[1] , **Ye Wang**[1] , **Xin Song**[1] , **Siwei Chen**[2] , **Han Yu**[1] and **Aiping Li**[1]
[1]National University of Defense Technology, Changsha, China
[2]Defense Innovation Institute, Beijing, China
{chenkai_, ye.wang, songxin, yuhan17, liaiping}@nudt.edu.cn, chensiwei1257@163.com

## Abstract

Temporal knowledge graph extrapolation has become a prominent area of study interest in recent years. Numerous methods for extrapolation have been put forth, mining query-relevant information from history to generate forecasts. However, existing approaches normally do not discriminate between causal and non-causal effects in reasoning; instead, they focus on analyzing the statistical correlation between the future events to be predicted and the historical data given, which may be deceptive and hinder the model's capacity to learn real causal information that actually affects the reasoning conclusions. To tackle it, we propose a novel approach called Causal Subhistory Identification (CSI), which focuses on extracting the causal subhistory for reasoning purposes from a large amount of historical data. CSI can improve the clarity and transparency of the reasoning process and more effectively convey the logic behind conclusions by giving priority to the causal subhistory and eliminating non-causal correlations. Extensive experiments demonstrate the remarkable potential of our CSI in the following aspects: superiority, improvement, explainability, and robustness.

## 1 Introduction

Knowledge graphs (KGs) are a powerful tool for representing and reasoning about structured knowledge [Song *et al.*, 2021; Zhao *et al.*, 2021; Yu *et al.*, 2024]. In recent years, there has been a growing interest in extending knowledge graphs to incorporate temporal information, leading to the development of temporal knowledge graphs (TKGs) [Trivedi *et al.*, 2017; Han *et al.*, 2020; Goel *et al.*, 2020]. TKGs capture not only static relationships between entities, but also time-varying relationships, thus enabling a more nuanced analysis of how knowledge evolves. With the temporal information included, a fact in TKGs are usually in the format of a quadruple, i.e., *(subject, relation, tail, timestamp)*.

Despite the ubiquity of TKGs, they remain far from being complete, primarily due to the limitations of our cognition. This has led to the emergence of TKG reasoning (TKGR), which involves inferring new facts based on existing knowledge. And there are generally two TKGR settings: interpolation [Xu *et al.*, 2020; Xu *et al.*, 2021; Chen *et al.*, 2022] and extrapolation [Jin *et al.*, 2020; Sun *et al.*, 2021; Liu *et al.*, 2022]. The interpolation seeks to complete missing facts within a specified range of timestamps. The extrapolation, on the other hand, predicts future events based on past knowledge. Our work aligns with extrapolation reasoning, which focuses on predicting future facts and presents a greater level of challenge [Li *et al.*, 2021; Liang *et al.*, 2023].

In recent years, various TKG extrapolation methods have been proposed, including those based on neural networks [Li *et al.*, 2021; Liang *et al.*, 2023] and logical rules [Liu *et al.*, 2022; Bai *et al.*, 2023], which have demonstrated impressive reasoning performance. These methods fundamentally rely on the statistical correlations between data, specifically the correlations between the input historical features and the queries to be answered, without fully considering the necessity of identifying the true evidence for reasoning. However, the statistical correlations between data encompass not only causal relationships but also numerous spurious or shortcut correlations [Pearl and others, 2000; Pearl, 2014]. These shortcuts are frequently the result of noisy features in the input data or biases in the data selection process [Sui *et al.*, 2022]. They can be misleading and can impede the model's ability to learn genuine causal evidence that truly influences the reasoning outcomes if not carefully accounted for in the analysis. Figure 1 depicts an example of a TKG that illustrates the spread of a specific epidemic virus among a population. When attempting to reason about *(Virus, infect, ?, 1-14)* (in dashed line), the reasoning model can arrive at the answer (entity *C*) through multiple paths:

a. $Virus \xrightarrow{infect,\ 1\text{-}10} A \xrightarrow{kiss,\ 1\text{-}11} B \xrightarrow{dine\ with,\ 1\text{-}13} C$,
b. $Virus \xrightarrow{infect,\ 1\text{-}10} A \xrightarrow{make\ a\ phone\ call,\ 1\text{-}12} C$,
c. $Virus \xrightarrow{infect,\ 1\text{-}12} D \xrightarrow{talk\ with,\ 1\text{-}9} C$.

Recent efforts [Knyazev *et al.*, 2019; Fan *et al.*, 2022; Wu *et al.*, 2022] reveal that reasoning methods are prone to exploiting shortcut features to make predictions and construct rationales. Although all three paths above lead to the answer, it is evident from a human perspective that only *Path a* is sufficiently compelling and can serve as evidence, while *Path b* and *Path c* are insufficient to genuinely support the reasoning outcome. Despite satisfying the statistical correlations, *Path*

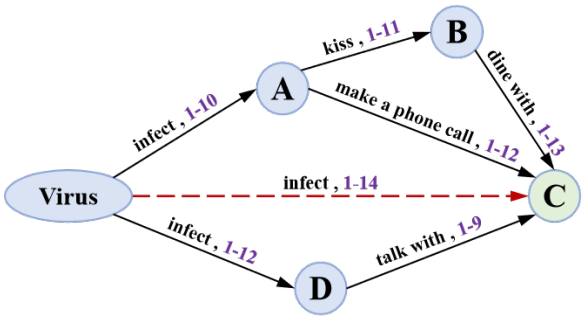



Figure 1: An example of a TKG illustrating the spread of a specific epidemic virus within a population.

*b* and *Path c* exhibit "spurious" or shortcut features, which do not genuinely contribute to the reasoning results. Therefore, it is crucial to use rigorous statistical analysis and experimental design to identify and control for these shortcuts and ensure that the conclusions drawn from data are based on solid causal evidence.

To this end, we aim to decouple historical information related to a query into two parts: 1) the **causal subhistory**, consisting of historical information that exhibits causal relationships with the query to be reasoned, and 2) the **shortcut subhistory**, consisting of non-causal but relevant features. We propose a **C**ausal **S**ubhistory **I**dentification (**CSI**) strategy, to maximize the causal effect between the causal subhistory and the predicted label, while reducing the confounding effect of the shortcut subhistory. To elaborate, we begin by utilizing an attention module to estimate both causal and shortcut features from the input TKG. Furthermore, leveraging causal theory [Pearl and others, 2000; Pearl, 2014], we parameterize the backdoor adjustments in the representation space, combining each causal estimate with various shortcut estimates and encouraging these combinations to maintain stable predictions. This approach promotes the invariant relationship between causal patterns and predictions, regardless of the variations and distributions in the shortcut components. By prioritizing the causal subhistory and excluding non-causal correlations, we can achieve greater clarity and transparency in our reasoning, enabling us to explain the rationale behind our conclusions more effectively. Extensive experiments demonstrate the promising capacity of our CSI from four aspects: superiority, improvement, explainability, and robustness.

We summarize our main contributions as follows:

1. We propose a novel causal subhistory identification strategy for extrapolation, to explain the rationale behind reasoning conclusions more effectively.

2. To the best of our knowledge, we are the first to consider the causality in TKG extrapolation. From a causal perspective, we analyze both the causal effects and non-causal effects that influence reasoning.

3. Comprehensive experiments show that our CSI outperforms SOTA extrapolation methods on the link prediction task, with decent interpretability and robustness.

## 2 Related Work

### 2.1 TKG Reasoning Methods

In recent years, a significant number of methods [Bordes *et al.*, 2013; Yang *et al.*, 2015; Dettmers *et al.*, 2018; Schlichtkrull *et al.*, 2018; Shang *et al.*, 2019] have been proposed for static KG reasoning. Furthermore, temporal knowledge graph (TKG) reasoning has also attracted attention, which can be categorized into two types: interpolation [Dasgupta *et al.*, 2018; García-Durán *et al.*, 2018; Xu *et al.*, 2019; Lacroix *et al.*, 2020; Jain *et al.*, 2020; Xu *et al.*, 2021] and extrapolation.

This work focuses on TKG extrapolation, which aims to predict facts in the future. Many methods based on neural networks have made significant contributions to improving the extrapolation reasoning performance, such as RE-NET [Jin *et al.*, 2020], RE-GCN [Li *et al.*, 2021], TANGO [Han *et al.*, 2021b], CyGNet [Zhu *et al.*, 2021], RE-GCN [Li *et al.*, 2021], HiSMatch [Li *et al.*, 2022b], and TiRGN [Li *et al.*, 2022a]. Recently, DaeMon [Dong *et al.*, 2023] adaptively models the temporal path information between the query subject and each object candidate across historical time periods. RPC [Liang *et al.*, 2023] employs relational GCN and gated recurrent units to mine relational correlations and periodic patterns from temporal facts. On the other hand, some other methods have been developed to enhance the reasoning explainability. TLogic [Liu *et al.*, 2022] automatically learn rules from data to improve reasoning performance and obtain explicit rules and explainable reasoning paths, while TITer [Sun *et al.*, 2021] explores possible reasoning paths using reinforcement learning. And xERTE [Han *et al.*, 2021a] provides explanations of the reasoning process and results through subgraph extraction and process tracing.

### 2.2 Causal Inference on Graphs

Causal theory [Pearl and others, 2000; Pearl, 2014; Morgan and Winship, 2015; Imbens and Rubin, 2015; Yao *et al.*, 2021] have provided researchers with new methods to design robust measurement approaches, discover hidden causal structures, and address data biases. Numerous studies have shown the benefits of incorporating causal relationships in various tasks involving graph neural networks. CFLP [Zhao *et al.*, 2022] leverages counterfactual links to augment data and improve link prediction. CAL [Sui *et al.*, 2022] intervenes on the representation of graph data to identify causal subgraphs involved in graph classification. CGI [Feng *et al.*, 2021] applies causal relationships to estimate the causal effects of local structures of nodes to assist in node classification. DIR [Wu *et al.*, 2022] applies causal reasoning to the interpretability of graphs and learns invariant principles through intervention on the training distribution. DisC [Fan *et al.*, 2022] and CMRL [Lee *et al.*, 2023] learn disentangled representations and combine each causal feature with various shortcut features to alleviate confounding effects. To the best of our knowledge, there is currently no existing work that utilizes causal theory to guide TKG reasoning.

# 3 Preliminaries

## 3.1 TKG Extrapolation

A Temporal Knowledge Graph (TKG) is a compilation of numerous temporal facts, denoted as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{F}\}$, where $\mathcal{E}$, $\mathcal{R}$, $\mathcal{T}$, and $\mathcal{F}$ represent the sets of entities, relations, timestamps, and facts, respectively. Each temporal fact within $\mathcal{G}$ is structured as a quadruple $(e_s, r, e_o, t) \in \mathcal{F}$, where a relation $r \in \mathcal{R}$ exists between a subject entity $e_s \in \mathcal{E}$ and an object entity $e_o \in \mathcal{E}$ at time $t \in \mathcal{T}$. Additionally, we employ inverse relations to enrich temporal facts and acquire $(e_o, r^{-1}, e_s, t)$ for each quadruple. Let $\mathcal{O} = (e_s, r, e_o, t)|t \in [t_0, T]$ represent the set of observable known facts, where $[t_0, T]$ denotes the accessible time interval. When presented with a query $Q = (e_q, r_q, ?, t_q)$, our TKG extrapolation endeavors to deduce the missing object entity given the other three elements. One prerequisite is $t_q > T$, indicating the initiation of extrapolation to infer future events based on past facts.

## 3.2 A Causal View on Extrapolation

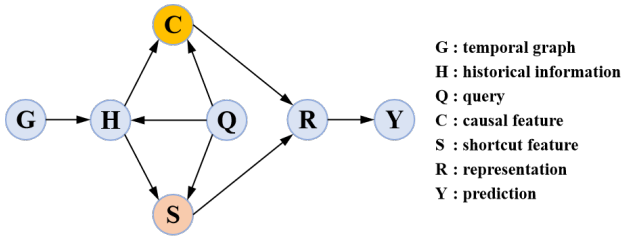



Figure 2: Structural causal model for TKG extrapolation.

We analyze the causal aspects of the process of TKG extrapolation from a causal perspective and depict a structural causal model (SCM) [Pearl and others, 2000] in Figure 2. The SCM illustrates the causal relationships among six variables: temporal graph $G$, historical information $H$, query $Q$, causal subhistory $C$, shortcut subhistory $S$, representation $R$, prediction $Y$. Specifically, the query can be further refined as $Q = (e_q, r_q, ?, t_q)$. And each arrow $\rightarrow$ in the SCM represents a causal-effect relationship, where cause $\rightarrow$ effect. Below are the explanations for each causal-effect relationship:

- $\boldsymbol{G \rightarrow H \leftarrow Q}$. The variable $H$ denotes the historical information that occurred before the query $Q$ (i.e., prior to $t_q$) and is derived from the temporal graph $G$. Generally, only the entities within a specific distance from $e_q$ and their connections within $G$ could be taken into account in $H$.
- $\boldsymbol{H \rightarrow C \leftarrow Q}$. The variable $C$ denotes the causal feature specific to the query $Q$, which truly reflects the intrinsic property of the historical history $H$, indicating the causal features that influence reasoning and their close correlation with the query.
- $\boldsymbol{H \rightarrow S \leftarrow Q}$. The variable $S$ represents the shortcut or redundant feature specific to the query $Q$, which is spuriously correlated with the prediction and usually caused by the data biases.
- $\boldsymbol{C \rightarrow R \leftarrow S}$. The variable $R$ denotes the representation of the missing entity to be predicted. $R$ is obtained by fusing information from message passing and aggregation operations on both the causal subhistory and the shortcut subhistory.
- $\boldsymbol{R \rightarrow Y}$. The ultimate goal of the representation learning is to predict the missing object entity given the query $Q = (e_q, r_q, ?, t_q)$. The classifier will make a prediction based learned representation $R$.

By analyzing the SCM, we can identify three backdoor paths that hinder the model's ability to learn the causal relationship between $C$ and $Y$: $C \leftarrow H \rightarrow S \rightarrow R \rightarrow Y$, $C \leftarrow Q \rightarrow S \rightarrow R \rightarrow Y$, and $C \leftarrow Q \rightarrow H \rightarrow S \rightarrow R \rightarrow Y$. Note that our reasoning is to infer the missing entitiy for a given query, meaning that the query $Q$ has been known to us. Thus, the latter two backdoor paths $C \leftarrow Q \rightarrow S \rightarrow R \rightarrow Y$ and $C \leftarrow Q \rightarrow H \rightarrow S \rightarrow R \rightarrow Y$ can be effectively blocked by conditioning on $Q$. However, the only remaining backdoor path, $C \leftarrow H \rightarrow S \rightarrow R \rightarrow Y$, will establish a spurious correlation between $C$ and $Y$, leading the model to make predictions based on $S$ rather than $C$. Therefore, it is crucial to block the backdoor path and enable the model to utilize the causal graph for prediction.

## 3.3 Backdoor Adjustment

We have realized that the shortcut feature $S$ has a confounding effect when analyzing the causal effect between $C$ and $Y$. Hence, safeguarding our reasoning model against the impact of $S$ is an essential challenge to face. Fortunately, causal theory [Pearl and others, 2000; Pearl, 2014] provides us with a flexible approach: we can exploit the do-calculus on the variable $C$ through backdoor adjustment, which involves stratifying the confounding element $S$. Ultimately, our backdoor adjustment can cut off the backdoor path $C \leftarrow H \rightarrow S \rightarrow R \rightarrow Y$, and obtain an intervened distribution $\hat{P}(Y|C, Q) = P(Y|do(C), Q)$.[1] And our backdoor adjustment can be formulated as follows:

$$\begin{aligned} P(Y|do(C), Q) &= \hat{P}(Y|C, Q) \\ &= \sum_{s \in \mathcal{T}_s} \hat{P}(Y|C, Q, s)\hat{P}(s|C, Q) \quad (Bayes\ Rule) \\ &= \sum_{s \in \mathcal{T}_s} \hat{P}(Y|C, Q, s)\hat{P}(s|Q) \quad (Independency) \\ &= \sum_{s \in \mathcal{T}_{sc}} P(Y|C, Q, s)P(s|Q)\,, \end{aligned} \tag{1}$$

where $\mathcal{T}_s$ denotes the confounder set, $P(Y|C, Q, s)$ denotes the conditional probability of the prediction given the causal feature $C$, the query $Q$, and the confounder $s$. $P(s|Q)$ is the conditional probability of the confounder given the query $Q$. Under the causal intervention, the causal variables $C$ and the confounder $S$ are independent, as denoted by $\hat{P}(s|C, Q) = \hat{P}(s|Q)$. Additionally, we have $\hat{P}(Y|C, Q, s) = P(Y|C, Q, s)$ and $\hat{P}(s|Q) = P(s|Q)$ because cutting off the backdoor path does not affect the con-

[1]The symbol $\hat{P}$ indicates that the probability is under the causal intervention.

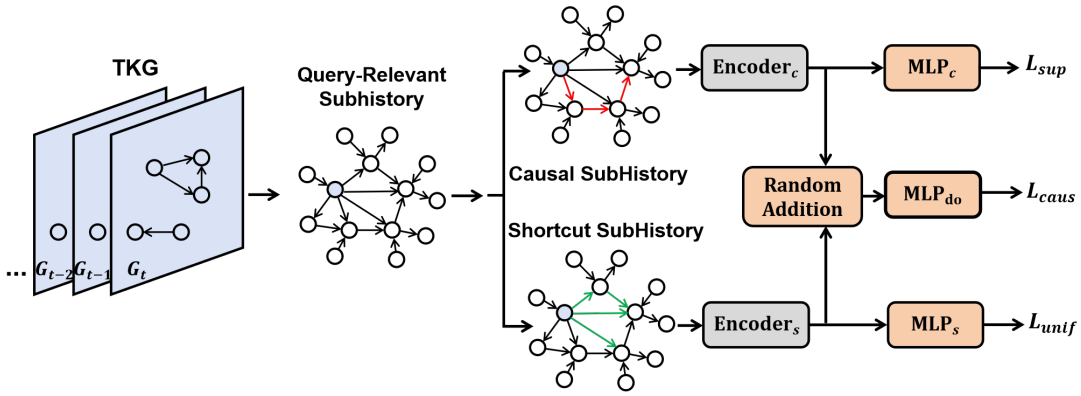



Figure 3: The overview of the CSI framework.

ditional distribution of $Y$ given $C, Q, s$, while the conditional distribution of $s$ given $Q$ remains unchanged.

Despite the effectiveness of backdoor adjustment in mitigating the confounder, there are three challenges in practical implementation: 1) The confounder set is often unobservable and difficult to obtain. 2) The dynamic evolution and multi-relational nature of TKGs present a significant challenge for direct manipulation of TKG data. 3) Due to the close relationship between the variable $C$ and the query $Q$, the process of identifying the causal feature from $H$ is inherently query-specific, adding complexity to the task. To address these issues, we propose an effective causal subhistory identification framework in Section 4.

# 4 Methodology

In this section, we provide a detailed description of our Causal Subhistory Identification (CSI) strategy for TKG extrapolation tasks, as illustrated in Figure 3.

## 4.1 Query-Relevant Subhistory Extraction

In the pursuit of efficiently predicting unknown future events based on historical information, our approach involves targeted extraction query-relevant background information from a vast amount of historical data, referred to as "query-relevant **subhistory**", to guide our reasoning. The concept of a subhistory can be essentially represented as a localized subgraph with time information. Given a query $Q = (e_q, r_q, ?, t_q)$, we closely focus on the query entity $e_q$ and consider the relevance with the query relation $r_q$ to obtain a subset of the historical data, i.e., the query-relevant subhistory, which is significant in addressing the query. The extraction process focuses on creating an entity-centric subgraph that is temporally and contextually relevant to the query entity, ensuring a focused and tailored representation of the historical data preceding the query time.

Consequently, the obtained subhistory is formed as a localized subgraph $\mathcal{G}_Q = \{\mathcal{E}_Q, \mathcal{R}_Q, \mathcal{T}_Q, \mathcal{F}_Q\}$, where $\mathcal{E}_Q \subseteq \mathcal{E}$, $\mathcal{R}_Q \subseteq \mathcal{R}$, $\mathcal{T}_Q \subseteq \mathcal{T}$, and $\mathcal{F}_Q \subseteq \mathcal{F}$ denote the sets of entities, relations, timestamps (prior to $t_q$) and facts related to the given query, respectively. The subhistory itself presents a targeted perspective on the query-relevant historical data, providing sufficient relevancy to support us in reasoning about the query. And all the pertinent historical information required to address the query has been effectively incorporated.

## 4.2 Disentangled Causal & Shortcut Subhistory

We have already discussed in Section 3.2 that the historical information relevant to the query essentially includes the causal feature and the shortcut feature. Thus, given the extracted query-relevant subhistory $\mathcal{G}_Q = \{\mathcal{E}_Q, \mathcal{R}_Q, \mathcal{T}_Q, \mathcal{F}_Q\}$, we further attempt to disentangle it into two parts: a **causal subhistory** $\mathcal{G}_c$ and a **shortcut subhistory** $\mathcal{G}_s$.

To achieve disentanglement, we formulate soft masks for the facts in $\mathcal{G}_Q$, where each element in the masks represents the attention score relevant to the task of interest, typically ranging between 0 and 1. Due to the mutual exclusivity of causal and non-causal features, $\mathcal{G}_c$ and $\mathcal{G}_s$ are complementary to each other. We assign the mask $M$ to $\mathcal{G}_c$ and the mask $\overline{M} = \mathbf{1} - M$ to $\mathcal{G}_s$, where $\mathbf{1}$ is a matrix filled with ones. Specifically, for the $i$-th fact in $\mathcal{F}_Q$, $M_{i|Q}$ represents the level of its involvement in $\mathcal{G}_c$ given $Q$, while $\overline{M}_{i|Q}$ represents the level of its involvement in $\mathcal{G}_s$ given $Q$.

In practice, to obtain quantifiable masks, we turn to the graph attention mechanism [Velickovic *et al.*, 2018] for assistance. We denote the $i$-th fact in $\mathcal{F}_Q$ by $(\mathcal{F}_Q)_i = (e_s, r, e_o, t)$. Within the extracted query-relevant $\mathcal{G}_Q$, we first need a message passing module, which enables nodes to exchange and aggregate information in a localized and adaptive manner, capturing the structural dependencies and patterns present in both $\mathcal{G}_c$ and $\mathcal{G}_s$. We focus on the message passing from the subject entity $e_s$ to the object entity $e_o$, and take the message value as a combination of the representations of the subject $e_s$, the relation $r$, and the time $t$:

$$Msg_i = h_{e_s} + h_r + h_t\,, \tag{2}$$

where we use the same dimension $d$ for $h_{e_s}$, $h_r$, and $h_t$.

**Causal Mask.** For the message $Msg_i$, we then make use of a multi-layer perception (MLP) to calculate the query-specific attention weight:

$$\alpha_{i|Q} = \mathrm{MLP}([h_{e_s}, h_r, h_{r_q}, h_t])\,, \tag{3}$$

where $h_{r_q}$ is the representation of the query relation $r_q$, and conditioning on $Q$ implies that the weight scores vary depending on different queries.

To map the obtained weight score into the range of (0,1), we employ a sigmoid function to get the causal mask:

$$M_{i|Q} = \sigma(\alpha_{i|Q}) \,. \tag{4}$$

By applying the sigmoid function, we ensure that the weights are transformed into a probabilistic representation, allowing for a more intuitive interpretation of their significance within the causal subhistory.

**Shortcut Mask.** We use a shortcut mask to represent statistical associations outside of causality in $\mathcal{F}_Q$, i.e., the shortcut or confounding features:

$$\overline{M}_{i|Q} = 1 - M_{i|Q} \,. \tag{5}$$

This choice of $M_{i|Q}$ or $\overline{M}_{i|Q}$ implies that the resulting values represent the probability of the corresponding fact being present in the causal subhistory or the shortcut subhistory. Ultimately, we can divide the full query-relevant subhistory $\mathcal{G}_Q$ into: the causal subhistory $\mathcal{G}_c = \{\mathcal{E}_Q, \mathcal{R}_Q, \mathcal{T}_Q, \mathcal{F}_Q \odot M\}$ and the shortcut subhistory $\mathcal{G}_s = \{\mathcal{E}_Q, \mathcal{R}_Q, \mathcal{T}_Q, \mathcal{F}_Q \odot \overline{M}\}$.

The soft masks, $M$ and $\overline{M}$, are associated with each fact through a weight score between 0 and 1, controlling the contribution of it in the information propagation and aggregation within the subhistory. Adhering to causal theory, the soft masks dynamically weight each fact according to its causal relationship strength to addressing the query. This weighting mechanism enables the encoder to handle causal relationships more effectively and highlight significant causal paths.

## 4.3 Learning for Prediction

With messages and the weights obtained, we adopt two GNN encoders to obtain entity representations in $\mathcal{G}_c$ and $\mathcal{G}_s$ conditioned on $Q$:

$$H_c|Q = \text{Encoder}_c(\mathcal{G}_c, Msg, M) \,, \tag{6}$$

$$H_s|Q = \text{Encoder}_s(\mathcal{G}_s, Msg, \overline{M}) \,, \tag{7}$$

where $H_c \in \mathbb{R}^{|\mathcal{E}_Q| \times d}$ and $H_s \in \mathbb{R}^{|\mathcal{E}_Q| \times d}$.

To obtain the likelihood of the entities, we utilize two MLP layers which act as powerful feature extractors from $\mathcal{G}_c$ and $\mathcal{G}_s$. Then, the sigmoid function is applied to each output to transform it into a probabilistic range between 0 and 1.

$$p_c|Q = \sigma(\text{MLP}_c(H_c|Q)) \,, \tag{8}$$

$$p_s|Q = \sigma(\text{MLP}_s(H_s|Q)) \,, \tag{9}$$

This transformation represents the likelihood of the entities serving as candidate answers for the given query, where values closer to 1 indicate a higher likelihood of the entities being relevant answers to the query.

**Loss for Causal Subhistory.** We expect the final prediction to be determined by the causal subhistory, and it is desirable for the causal subhistory to reflect as much information about the predicted label as possible. With labels available, we employ a multi-class log-loss function to obtain a supervised loss and train the neural networks for our TKG Reasoning, which has been proven effective [Lacroix *et al.*, 2018; Zhang and Yao, 2022]:

$$\mathcal{L}_{sup} = \sum_{(Q,e_a)\in\mathcal{T}_{\text{train}}} \left( -p_c(e_a)|Q + \log \left( \sum_{\forall e \in \mathcal{E}} e^{p_c(e)|Q} \right) \right) , \tag{10}$$

where $e_a$ is the answer entity for the query Q in the training set $\mathcal{T}_{\text{train}}$, and $p_c(e)|Q$ denotes the likelihood of the entity $e$ serves as a candidate answer for the given query $Q$.

**Loss for Shortcut Subhistory.** Simultaneously, we encourage the shortcut subhistory to contain as little information as possible that is related to the predicted label, in other words, to depersonalize the shortcut subhistory. And we expect it to approximate a uniform distribution:

$$\mathcal{L}_{unif} = \text{KL}(y_{unif}, p_s|Q) \,, \tag{11}$$

where KL denotes the KL-Divergence, $y_{unif}$ represents the uniform distribution. The approximation of a uniform distribution further ensures diverse and evenly spread contextual cues in the shortcut subhistory, reducing the risk of being influenced by irrelevant factors and improving the model's generalization capability.

By leveraging $\mathcal{L}_{pred}$ and $\mathcal{L}_{unif}$, we successfully accomplish a disentanglement between the causal subhistory and the shortcut subhistory. These loss functions enable us to differentiate between these two subhistories by promoting the causal subhistory to capture significant contextual cues that are pertinent to the predicted label. Simultaneously, they encourage the shortcut subhistory to encompass impartial and varied cues that are unrelated to the label.

## 4.4 Causal Intervention

As analyzed in Section 3.3, the key to backdoor adjustment lies in intervening on the causal variable $C$. However, directly intervening on data-level is not practical due to TKGs' dynamic evolution and multi-relational nature. Therefore, we consider implicitly intervening on representation-level, by generating the counterfactual unconfounded samples in embedding space. More specifically, we randomly permute shortcut representations in each mini-batch and obtain $H|Q_{(do)} = H_c|Q + \widehat{H_s}|Q$, where $\widehat{H_s}|Q$ denotes the randomly permuted shortcut representations of $H_s|Q$. We also prepare a MLP layer $\text{MLP}_{do}$ and the sigmoid function for the intervened $H|Q_{(do)}$: $p_{do}|Q = \sigma(\text{MLP}_{do}(H|Q_{(do)}))$. With the generated unconfounded samples, we utilize the following loss function guided by the backdoor adjustment:

$$\mathcal{L}_{caus} = \sum_{(Q,e_a)\in\mathcal{T}_{\text{train}}} \left( -p_{do}(e_a)|Q + \log \left( \sum_{\forall e \in \mathcal{E}} e^{p_{do}(e)|Q} \right) \right) , \tag{12}$$

which leverages causal features to ensure the predictions remain invariant and stable across diverse contexts.

Together with the losses for two subhistories, our total loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{unif} + \lambda_2 \mathcal{L}_{caus} \,, \tag{13}$$

where $\lambda_1$ and $\lambda_2$ are hyper-parameters that control the strength of disentanglement and causal intervention.

# 5 Experiments

In this section, we assess the effectiveness of our proposed CSI on four key datasets, aiming to answer four pivotal questions through experimental results and analysis:

| Model | GDELT | | | | ICEWS14* | | | | WIKI | | | | ICEWS05-15 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | H@1 | H@3 | H@10 | MRR | H@1 | H@3 | H@10 | MRR | H@1 | H@3 | H@10 | MRR | H@1 | H@3 | H@10 |
| RE-NET [2020] | 19.6 | 12.4 | 21.0 | 34.0 | 38.3 | 28.7 | 41.3 | 54.5 | 49.7 | 46.9 | 51.2 | 43.5 | 43.3 | 33.4 | 47.8 | 63.1 |
| TANGO [2021b] | 19.2 | 12.2 | 20.4 | 32.8 | 26.3 | 17.3 | 29.1 | 44.2 | 50.4 | 48.5 | 51.5 | 53.6 | 42.9 | 32.7 | 48.1 | 62.3 |
| CyGNet [2021] | 18.5 | 11.5 | 19.6 | 32.0 | 32.7 | 23.7 | 36.3 | 50.7 | 58.8 | 47.9 | 66.4 | 78.7 | 36.8 | 26.6 | 41.6 | 56.2 |
| TITer [2021] | 20.2 | 14.1 | 22.2 | 31.2 | 41.7 | 32.7 | 46.5 | 58.4 | 73.9 | 71.7 | 75.4 | 77.0 | 47.7 | 38.0 | 52.9 | 65.8 |
| RE-GCN [2021] | 19.8 | 12.5 | 21.0 | 34.0 | 41.8 | 31.6 | 46.7 | 61.5 | 78.5 | 74.5 | 81.6 | 84.7 | 48.0 | 37.3 | 53.9 | 68.3 |
| xERTE [2021a] | 18.9 | 12.3 | 20.1 | 30.3 | 40.8 | 32.7 | 45.7 | 57.3 | 71.1 | 68.1 | 76.1 | 79.0 | 46.6 | 37.8 | 52.3 | 63.9 |
| TiRGN [2022a] | 21.7 | 13.6 | 23.3 | 37.6 | 45.1 | 34.4 | 51.3 | 65.0 | 81.7 | 77.8 | 85.1 | 87.1 | 50.0 | 39.3 | 56.1 | 70.7 |
| HiSMatch [2022b] | 22.0 | 14.5 | 23.8 | 36.6 | 45.8 | 35.8 | 50.8 | 65.1 | 78.1 | 73.9 | 81.32 | 84.7 | 52.9 | 42.0 | 59.1 | 73.3 |
| TLogic [2022] | 19.8 | 12.2 | 21.7 | 35.6 | 43.0 | 33.6 | 48.3 | 61.2 | 79.7 | 75.4 | 81.9 | 85.2 | 47.0 | 36.2 | 53.1 | 67.4 |
| DaeMon [2023] | 20.7 | 13.7 | 22.5 | 34.2 | - | - | - | - | 82.4 | 78.3 | 86.0 | 88.0 | - | - | - | - |
| RPC [2023] | 22.4 | 14.4 | 24.4 | 38.3 | 44.6 | 34.9 | 49.8 | 65.1 | 81.2 | 76.3 | 85.4 | 88.7 | 51.4 | 39.9 | 57.0 | 71.8 |
| CSI (ours) | **25.6** | **18.2** | **28.5** | **40.7** | **47.5** | **37.4** | **52.8** | **67.0** | **83.7** | **80.4** | **87.9** | **89.6** | **53.3** | **42.5** | **59.2** | **73.6** |

Table 1: Performance (in percentage) for link prediction on four benchmarks with time-aware metrics.

| Datasets | GDELT | ICEWS14* | WIKI | ICEWS05-15 |
|---|---|---|---|---|
| Entities | 7,691 | 7,128 | 12,554 | 10,094 |
| Relations | 240 | 230 | 24 | 251 |
| Train | 1734,399 | 63,685 | 539,286 | 368,868 |
| Validation | 238,765 | 13,823 | 67,538 | 46,302 |
| Test | 305,241 | 13,222 | 63,110 | 46,159 |
| Time granularity | 15 mins | 24 hours | 1 year | 24 hours |
| Time Stamps | 2975 | 365 | 232 | 4017 |

Table 2: Statistics of the datasets.

**Q1: Superiority.** How does our CSI perform compared to existing methods?

**Q2: Improvement.** How does each component of CSI contribute to the performance improvement?

**Q3: Explainability.** How does CSI enhance the explainability of reasoning?

**Q4: Robustness.** How does the robustness of CSI perform under data sparsity & data noise?

## 5.1 Experimental Setup

**Benchmark Datasets**

Four TKGR benchmark datasets are leveraged to evaluate our CSI, including ICEWS14* [Han *et al.*, 2021a], ICEWS05-15 [García-Durán *et al.*, 2018], WIKI [Leblay and Chekol, 2018], and GDELT [Jin *et al.*, 2020]. Details of the four datasets we use are shown in Table 2. ICEWS14* and ICEWS05-15 are two subsets of the large-scale event-based database, Integrated Crisis Early Warning System [Lautenschlager *et al.*, 2015]. The ICEWS14* dataset encompass events that took place in 2014, while the ICEWS05-15 dataset includes events that occurred between 2005 and 2015. We take drop the month and date information of WIKI here for ease of processing, and obtain the same year-level granularity as [Jin *et al.*, 2020]. GDELT is extracted from the global database of events, language, and tone [Leetaru and Schrodt, 2013], which has a fine-grained time granularity of 15 minutes.

**Evaluation Protocol**

The evaluation for TKG extrapolation involves the adoption of a link prediction task. This task focuses on inferring incomplete time-wise facts that contain a missing entity, represented as either $(e_s, r, ?, t)$ or $(?, r, e_o, t)$. We use the ground truths for extrapolation, as is the case with many previous methods [Jin *et al.*, 2020; Li *et al.*, 2021]. Specifically, for all of the training, validation and testing, we predict future events assuming ground truths of the preceding events are given at inference time [Han *et al.*, 2021a]. And we use the time-wise filtered setting [Xu *et al.*, 2019; Goel *et al.*, 2020] to report the experimental results. The performance is reported on the standard evaluation metrics: the proportion of correct triples ranked in the top 1, 3 and 10 (Hits@1, Hits@3, and Hits@10), and Mean Reciprocal Rank (MRR). All the metrics are the higher the better. For all experiments, we report averaged results across 5 runs, and we omit the variance as it is generally low.

**Baselines**

We compare with eleven up-to-date TKG extrapolation baseline methods, including RE-NET [Jin *et al.*, 2020], TANGO [Han *et al.*, 2021b], CyGNet [Zhu *et al.*, 2021], TITer [Sun *et al.*, 2021], RE-GCN [Li *et al.*, 2021], xERTE [Han *et al.*, 2021a], TiRGN [Li *et al.*, 2022a], HiSMatch [Li *et al.*, 2022b], TLogic [Liu *et al.*, 2022], DaeMon [Dong *et al.*, 2023], RPC [Liang *et al.*, 2023].

## 5.2 Performance Comparison (RQ1)

The experimental results obtained from four different extrapolation datasets are presented in Table 1. The datasets employed in our evaluation exhibit significant differences in scale, number of entities, and number of relations, as shown in Table 2. Our findings demonstrate the effectiveness of our proposed method in performing efficient TKG extrapolation of varying sizes and complexity levels. One notable observation is that our method outperforms all baseline approaches across all four datasets. This finding highlights the superiority of our reasoning performance and indicates the potential of our proposed model to tackle complex reasoning tasks over knowledge graphs. Our approach is particularly promising for addressing real-world problems involving vast amounts of semantic data from multiple sources, where accurate reasoning is crucial for making informed decisions.

## 5.3 Ablation Study (RQ2)

According to Equation 13, $\lambda_1$ and $\lambda_2$ control the strength of disentanglement and causal intervention, respectively. We then conduct ablation studies on two datasets, namely

| Model | ICEWS14* | | ICEWS05-15 | |
|---|---|---|---|---|
| | MRR | H@10 | MRR | H@10 |
| CSI | 47.5 | 67.0 | 53.3 | 73.6 |
| -$\lambda_1$ | 45.9 | 65.8 | 52.1 | 71.8 |
| -$\lambda_2$ | 46.7 | 66.4 | 53.0 | 73.2 |
| -$\lambda_1$-$\lambda_2$ | 45.2 | 65.5 | 51.5 | 71.0 |

Table 3: Ablation studies on ICEWS14* and ICEWS05-15.

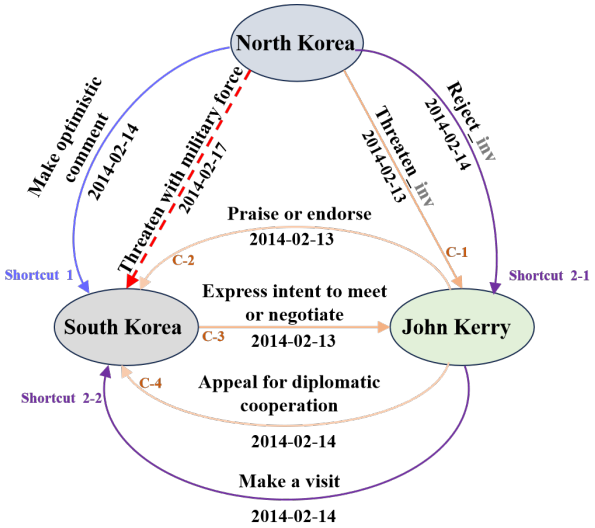



Figure 4: Case study of TKG extrapolation on ICEWS14*, where the query is indicated by the red dashed line, and "_inv" denotes the inverse relation.

ICEWS14* and ICEWS05-15, to investigate the impact of individual components on the performance of our proposed CSI in extrapolation. The MRR and Hits@10 performances are shown in Table 3, where three sub-models are compared, including (1) the original CSI, (2) CSI without depersonalizing shortcut subhistory, denoted as "-$\lambda_1$", (3) CSI without representation-level intervene, denoted as "-$\lambda_2$", (4) CSI without both depersonalizing shortcut subhistory and representation-level intervention, denoted as "-$\lambda_1$-$\lambda_2$". By comparing the results, we can find out that "-$\lambda_1$", "-$\lambda_2$", and "-$\lambda_1$-$\lambda_2$" all lead to a decrease in the performance of our CSI, indicating that both the shortcut term and the causal intervention term contribute to the improvement of our results. Additionally, "-$\lambda_1$" has a greater reduction in performance than "-$\lambda_2$", suggesting that the shortcut term has a larger impact on reasoning performance.

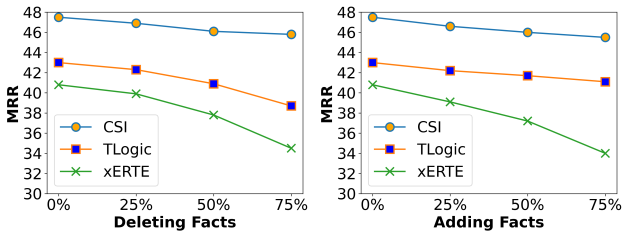
### 5.4 Case Study: Reasoning Explainability (RQ3)

Case studies are taken to illustrate the advantages of our CSI in reasoning explainability. Figure 4 depicts a reasoning diagram for the query *(North Korea, Threaten with military force, ?, 2014-02-17)*, including a causal clue (C-1 → C-2 → C-3 → C-4) and two spurious paths (Shortcut 1 and Shortcut 2) generated by shortcut features. It is evident that while it is possible to reach the answer from the query entity through these spurious paths, they lack explanatory power and fail to provide an adequate justification for compelling reasoning to resolve the query. The scores for these paths, obtained by combining the fact scores, are displayed in Table 4, with and without the utilization of our proposed CSI. It is noteworthy that without the employment of CSI, the reasoning model tends to prioritize spurious paths, whereas our CSI effectively mitigates the influence of such paths.

| CSI | Casual Clue | Spurious Clue | |
|---|---|---|---|
| | | Shortcut 1 | Shortcut 2 |
| W/ | 0.22 | 0.04 | 0.13 |
| W/O | 0.09 | 0.12 | 0.10 |

Table 4: Path scores for causal clue and spurious clues.

### 5.5 Robustness Evaluation (RQ4)




(a) Data sparsity (b) Data noise

Figure 5: Robustness evaluation on ICEWS14*.

To evaluate the robustness of CSI, we experiment with two scenarios: data sparsity and data noise, by deleting or adding facts to ICEWS14*. Specifically, we randomly remove (if facts exist) or add (if no such facts) 25%, 50%, 75% facts in the training set. Figure 5 shows a performance comparison across varying degrees of data sparsity and data noise, where we compare our CSI with two baselines: the rule-based TLogic [Liu *et al.*, 2022] and the GNN-based xERTE [Han *et al.*, 2021a]. Compared to them, our CSI shows remarkable performance across a broad spectrum of sparsity and noise levels. In particular, even when the level of sparsity or noise comes to 75%, our CSI continues to achieve a high MRR value exceeding 45. This highlights its exceptional robustness and ability to handle sparse data and noisy data effectively.

## 6 Conclusion

In this paper, we delve into the intricacies of Temporal Knowledge Graph (TKG) extrapolation through the lens of causality, introducing a novel approach known as the Causal Subgraph Interventions (CSI) strategy. By prioritizing the causal subhistory and excluding non-causal correlations, we achieve greater clarity and transparency in our reasoning, which enables us to explain the rationale behind our conclusions more effectively. Extensive experiments demonstrate the promising capacity of our CSI from four aspects, i.e., superiority, improvement, explainability, and robustness.

## Acknowledgments

This work is particularly supported by the National Key Research and Development Program of China (No. 2022YFB31040103), and the National Natural Science Foundation of China (No. 62302507).

## Contribution Statement

All authors conceived the paper's methodology and experiments. Kai Chen and Ye Wang contributed equally to this work and should be considered co-first authors. Both Han Yu and Aiping Li are corresponding authors.

## References

[Bai *et al.*, 2023] Luyi Bai, Wenting Yu, Die Chai, Wenjun Zhao, and Mingzhuo Chen. Temporal knowledge graphs reasoning with iterative guidance by temporal logical rules. *Information Sciences*, 621:22–35, 2023.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[Chen *et al.*, 2022] Kai Chen, Ye Wang, Yitong Li, and Aiping Li. Rotateqvs: Representing temporal information as rotations in quaternion vector space for temporal knowledge graph completion. In *ACL 2022*, 2022.

[Dasgupta *et al.*, 2018] Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. Hyte: Hyperplane-based temporally aware knowledge graph embedding. In *EMNLP 2018*, pages 2001–2011, 2018.

[Dettmers *et al.*, 2018] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *AAAI 2018*, 2018.

[Dong *et al.*, 2023] Hao Dong, Zhiyuan Ning, Pengyang Wang, Ziyue Qiao, Pengfei Wang, Yuanchun Zhou, and Yanjie Fu. Adaptive path-memory network for temporal knowledge graph reasoning. In *IJCAI 2023*, 2023.

[Fan *et al.*, 2022] Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. In *NeurIPS 2022*, 2022.

[Feng *et al.*, 2021] Fuli Feng, Weiran Huang, Xiangnan He, Xin Xin, Qifan Wang, and Tat-Seng Chua. Should graph convolution trust neighbors? A simple causal inference method. In *SIGIR 2021*, 2021.

[García-Durán *et al.*, 2018] Alberto García-Durán, Sebastijan Dumancic, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *EMNLP 2018*, 2018.

[Goel *et al.*, 2020] Rishab Goel, Seyed Mehran Kazemi, Marcus A. Brubaker, and Pascal Poupart. Diachronic embedding for temporal knowledge graph completion. In *AAAI 2020*, 2020.

[Han *et al.*, 2020] Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Dyernie: Dynamic evolution of riemannian manifold embeddings for temporal knowledge graph completion. In *EMNLP 2020*, 2020.

[Han *et al.*, 2021a] Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *ICLR 2021*, 2021.

[Han *et al.*, 2021b] Zhen Han, Zifeng Ding, Yunpu Ma, Yujia Gu, and Volker Tresp. Learning neural ordinary equations for forecasting future links on temporal knowledge graphs. In *EMNLP 2021*, 2021.

[Imbens and Rubin, 2015] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[Jain *et al.*, 2020] Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. Temporal knowledge base completion: New algorithms and evaluation protocols. In *EMNLP 2020*, 2020.

[Jin *et al.*, 2020] Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inference over temporal knowledge graphs. In *EMNLP 2020*, pages 6669–6683, 2020.

[Knyazev *et al.*, 2019] Boris Knyazev, Graham W. Taylor, and Mohamed R. Amer. Understanding attention and generalization in graph neural networks. In *NeurIPS 2019*, 2019.

[Lacroix *et al.*, 2018] Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2869–2878, 2018.

[Lacroix *et al.*, 2020] Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. Tensor decompositions for temporal knowledge base completion. In *ICLR 2020*, 2020.

[Lautenschlager *et al.*, 2015] Jennifer Lautenschlager, Steve Shellman, and Michael Ward. Icews events and aggregations. *Harvard Dataverse*, 3, 2015.

[Leblay and Chekol, 2018] Julien Leblay and Melisachew Wudage Chekol. Deriving validity time in knowledge graph. In *Companion Proceedings of the The Web Conference 2018*, pages 1771–1776, 2018.

[Lee *et al.*, 2023] Namkyeong Lee, Kanghoon Yoon, Gyoung S. Na, Sein Kim, and Chanyoung Park. Shift-robust molecular relational learning with causal substructure. In *SIGKDD 2023*, 2023.

[Leetaru and Schrodt, 2013] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer, 2013.

[Li *et al.*, 2021] Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. Temporal knowledge graph reasoning based

on evolutional representation learning. In *SIGIR 2021*, 2021.

[Li *et al.*, 2022a] Yujia Li, Shiliang Sun, and Jing Zhao. Tirgn: Time-guided recurrent graph network with local-global historical patterns for temporal knowledge graph reasoning. In *IJCAI 2022*, 2022.

[Li *et al.*, 2022b] Zixuan Li, Zhongni Hou, Saiping Guan, Xiaolong Jin, Weihua Peng, Long Bai, Yajuan Lyu, Wei Li, Jiafeng Guo, and Xueqi Cheng. Hismatch: Historical structure matching based temporal knowledge graph reasoning. In *EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7328–7338, 2022.

[Liang *et al.*, 2023] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In *SIGIR 2023*, 2023.

[Liu *et al.*, 2022] Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *AAAI 2022*, 2022.

[Morgan and Winship, 2015] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[Pearl and others, 2000] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

[Pearl, 2014] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

[Schlichtkrull *et al.*, 2018] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC 2018*, 2018.

[Shang *et al.*, 2019] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In *AAAI 2019*, pages 3060–3067, 2019.

[Song *et al.*, 2021] Yichen Song, Aiping Li, Hongkui Tu, Kai Chen, and Chenchen Li. A novel encoder-decoder knowledge graph completion model for robot brain. *Frontiers Neurorobotics*, 15:674428, 2021.

[Sui *et al.*, 2022] Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. Causal attention for interpretable and generalizable graph classification. In *KDD 2022*, pages 1696–1705, 2022.

[Sun *et al.*, 2021] Haohai Sun, Jialun Zhong, Yunpu Ma, Zhen Han, and Kun He. Timetraveler: Reinforcement learning for temporal knowledge graph forecasting. In *EMNLP 2021*, pages 8306–8319, 2021.

[Trivedi *et al.*, 2017] Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *ICML 2017*, 2017.

[Velickovic *et al.*, 2018] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR 2018*, 2018.

[Wu *et al.*, 2022] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR 2022*, 2022.

[Xu *et al.*, 2019] Chengjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Jens Lehmann, and Hamed Shariat Yazdi. Temporal knowledge graph embedding model based on additive time series decomposition. *CoRR*, abs/1911.07893, 2019.

[Xu *et al.*, 2020] Chengjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Hamed Shariat Yazdi, and Jens Lehmann. TeRo: A time-aware knowledge graph embedding via temporal rotation. In *COLING 2020*, 2020.

[Xu *et al.*, 2021] Chengjin Xu, Yung-Yu Chen, Mojtaba Nayyeri, and Jens Lehmann. Temporal knowledge graph completion using a linear temporal regularizer and multivector embeddings. In *NAACL-HLT 2021*, 2021.

[Yang *et al.*, 2015] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR 2015*, 2015.

[Yao *et al.*, 2021] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5):74:1–74:46, 2021.

[Yu *et al.*, 2024] Han Yu, Ziniu Liu, Hongkui Tu, Kai Chen, and Aiping Li. Generalizable inductive relation prediction with causal subgraph. *World Wide Web*, 27(3):1–19, 2024.

[Zhang and Yao, 2022] Yongqi Zhang and Quanming Yao. Knowledge graph reasoning with relational digraph. In *WWW 2022*, pages 912–924, 2022.

[Zhao *et al.*, 2021] Xiaojuan Zhao, Yan Jia, Aiping Li, Rong Jiang, Kai Chen, and Ye Wang. Target relational attention-oriented knowledge graph reasoning. *Neurocomputing*, 461:577–586, 2021.

[Zhao *et al.*, 2022] Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, and Meng Jiang. Learning from counterfactual links for link prediction. In *ICML 2022*, 2022.

[Zhu *et al.*, 2021] Cunchao Zhu, Muhao Chen, Changjun Fan, Guangquan Cheng, and Yan Zhang. Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. In *AAAI 2021*, 2021.