
Towards Foundation Model on Temporal Knowledge Graph Reasoning

Jiaxin Pan¹ Mojtaba Nayyeri¹ Osama Mohammed¹ Daniel Hernández¹

Rongchuan Zhang¹ Cheng Cheng¹ Steffen Staab^{1,2}

¹University of Stuttgart ²University of Southampton

{jiaxin.pan, mojtaba.nayyeri, osama.mohammed, daniel.hernandez}@ki.uni-stuttgart.de
{st191486, st180913}@stud.uni-stuttgart.de
{steffen.staab}@ki.uni-stuttgart.de

Abstract

Temporal Knowledge Graphs (TKGs) store temporal facts with quadruple formats (s, p, o, τ) . Existing Temporal Knowledge Graph Embedding (TKGE) models perform link prediction tasks in transductive or semi-inductive settings, which means the entities, relations, and temporal information in the test graph are fully or partially observed during training. Such reliance on seen elements during inference limits the models' ability to transfer to new domains and generalize to real-world scenarios. A central limitation is the difficulty in learning representations for entities, relations, and timestamps that are transferable and not tied to dataset-specific vocabularies. To overcome these limitations, we introduce the first fully-inductive approach to temporal knowledge graph link prediction. Our model employs sinusoidal positional encodings to capture fine-grained temporal patterns and generates adaptive entity and relation representations using message passing conditioned on both local and global temporal contexts. Our model design is agnostic to temporal granularity and time span, effectively addressing temporal discrepancies across TKGs and facilitating time-aware structural information transfer. As a pretrained, scalable, and transferable model, POSTRA demonstrates strong zero-shot performance on unseen temporal knowledge graphs, effectively generalizing to novel entities, relations, and timestamps. Extensive theoretical analysis and empirical results show that a single pretrained model can improve zero-shot performance on various inductive temporal reasoning scenarios, marking a significant step toward a foundation model for temporal KGs.

1 Introduction

Temporal Knowledge Graphs (TKGs) extend static knowledge graphs with temporal information. In TKGs, temporal facts are represented by the quadruple format (s, p, o, τ) , where s, p, o and τ denote head entity, relation name, tail entity and timestamp, respectively. Temporal Knowledge Graphs (TKGs) have been researched in various fields, including question answering [Saxena et al., 2021], data integration [Ao et al., 2022], and entity alignment [Cai et al., 2024].

Link prediction is a crucial task in temporal knowledge graphs (TKGs), which involves predicting missing entities in temporal facts. Given a query of the form $(?, p, o, \tau)$ or $(s, p, ?, \tau)$, the objective is to infer head or tail entity at a specific given timestamp τ . This task is essential for learning effective temporal embeddings that support various downstream applications. However, existing temporal knowledge graph embedding methods that follow transductive or semi-inductive settings suffer from transferability issues. As shown in Fig. 1 (a), the transductive setting (temporal knowledge graph interpolation) shares all the entities, relations, and timestamps information in both training and testing.

Fig. 1 (b) depicts the semi-inductive setting (temporal knowledge graph extrapolation or forecasting), where the timestamp information in the test graph is not available during training. However, the entities and/or relations remain consistent between the training and test graphs.

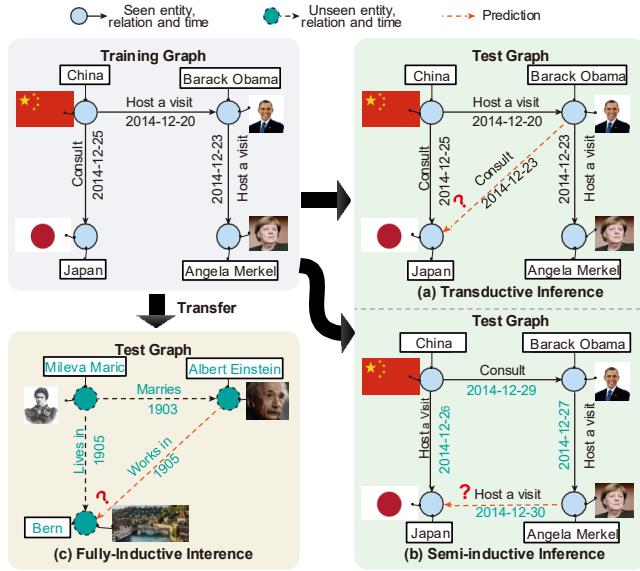


Figure 1: Subfigure (a) shows the *transductive* setting, where the test graph contains only entities, relation names, and timestamps that have already been seen during training. Subfigure (b) demonstrates the *semi-inductive* setting, where all entities and relation names in the test graph are present in the training graph, but the timestamps (e.g., 2014-12-30) are strictly later than those observed during training. Subfigure (c) illustrates the proposed *fully-inductive* setting, where the test graph includes unseen entities (e.g., Mileva Marić), relation names (e.g., lives in), and timestamps (e.g., 1905).

Zhang et al., 2022] often rely on dataset-specific trained time embeddings to transfer learned temporal information across TKGs with different granularities and time spans. To address this, we leverage the transferability of relative temporal ordering between connected facts along temporal paths. We organize a TKG as a sequence of temporal snapshots, where each snapshot contains all facts sharing the same timestamp. We focus on the relative ordering of quadruples in different snapshots, rather than the specific timestamp values, and encode this ordering using sinusoidal positional encodings as time series embeddings. For example, in Figure 2, although the temporal facts in two TKGs below have different timestamps, the relative temporal ordering difference between connected temporal facts is the same as shown in the relation graphs above. Therefore, this representation captures the temporal dependencies between facts in a way that is invariant to the underlying time units or time spans and transferable to new domains, effectively bypassing the temporal discrepancies challenge. By integrating these relative temporal signals into the message passing process of relation encoder and quadruple encoder, the model can learn temporal patterns without relying on dataset-specific granularity or time span.

To address the second challenge of temporal-aware structural information transfer, we adopt four core relation interaction types—head-head, head-tail, tail-head, and tail-tail—that are independent of dataset-specific relation names[Galkin et al., 2024, Lee et al., 2023b] and introduce a temporal-aware message passing mechanism. By constructing entity and relation embeddings as functions conditioned on these intrinsic interactions, a pre-trained model can inductively generalize to any knowledge graph through the message-passing process, even when encountering unseen entities and relations. To incorporate temporal dynamics, POSTRA injects temporal positional embeddings into the message-passing process, enabling the resulting entity and relation representations to capture

As existing TKG embedding models rely on the entity, relation vocabularies and used timestamps to train specific embeddings for prediction, they need to be retrained whenever a new TKG is introduced. This limitation hinders their generalization to real-world scenarios when new entities, relations, and timestamps emerge (Fig. 1 c). To overcome the generalization limitation in TKG models, we present a fully inductive TKG learning framework, **POSTRA**, a model for positional time-aware, and transferable inductive TKG representations. Specifically, we address two key challenges: 1) Temporal discrepancies across TKGs and 2) Temporal-aware structural information transfer across TKGs.

Starting with the first challenge, different TKGs utilize varying time units (e.g., minute/day/month) and cover diverse temporal spans (e.g., 1 month/100 years) depending on the frequency of temporal facts. Existing TKG embedding models[Lacroix et al., 2020,

time-dependent patterns, as illustrated in Figure 2. Moreover, to differentiate between queries that share the same entities and relation names but occur at different times (e.g., (s, p, o, τ) v.s. (s, p, o, τ')), POSTRA adopts a dual training strategy that separately learns local and global temporal structures. Global training aggregates information across the entire training graph, capturing long-term temporal dependencies, while local training restricts updates to a fixed temporal window around each target query, focusing on short-term temporal patterns. By combining these two complementary signals, POSTRA effectively adapts to TKGs with diverse temporal dynamics and varying time spans.

Different from existing TKGE models, our model does not rely on any dataset-specific entity, relation, or time embeddings (dataset-specific vocabulary embedding) and can generalize to any new TKG. While this allows the pretrained model to perform zero-shot inference across diverse TKGs, it also remains compatible with transductive and semi-inductive TKG link prediction tasks, providing flexibility across different task settings. Our model’s strong performance in transferability and cross-domain generalization across diverse datasets underscores its effectiveness and marks a key advance toward a foundation model for temporal knowledge graphs.

2 Task Formulation

Let V be a finite set of *entities*, R be a finite set of *relation names*, and $T = (\tau_i)_{i=1}^{|T|}$ be a finite ordered set of *timestamps*. A *temporal knowledge graph* is a quad $G = (V, R, T, Q)$, where $Q \subseteq V \times R \times V \times T$. The elements of Q are called *temporal facts* in quadruple form. Given a positive natural number i , with $i \leq |T|$, the i -*snapshot* of G , denoted G_i , is the subgraph of G that contains all temporal facts whose timestamp is the i -th element of T . We write τ_i for the timestamp at snapshot G_i . Intuitively, a temporal fact describes a relationship between two entities at a determined timestamp, and a snapshot describes all relationships that occur simultaneously at a determined timestamp.

Link Prediction. We aim at link prediction, the fundamental task for TKG reasoning which predicts missing entities of queries $(s, p, ?, \tau)$ or $(?, p, o, \tau)$. For this task, we assume: (i) one *training graph* $G_{train} = (V_{train}, R_{train}, T_{train}, Q_{train})$. (ii) one *inference graph* $G_{inf} = (V_{inf}, R_{inf}, T_{inf}, Q_{inf})$. We divide Q_{inf} into three pairwise disjoint sets, such that $Q_{inf} = Q_0 \cup Q_{valid} \cup Q_{test}$. $G_0 = (V_0, R_0, T_0, Q_0)$, $G_{valid} = (V_{valid}, R_{valid}, T_{valid}, Q_{valid})$, $G_{test} = (V_{test}, R_{test}, T_{test}, Q_{test})$ refers to the *observed graph*, *validation graph* and *test graph*, respectively.

At training time, we use G_{train} to train the model to predict Q_{train} . At inference, we use Q_0 to compute the embeddings and tune hyperparameters based on the model’s performance on Q_{valid} . Finally, we test the model’s performance on Q_{test} .

Transductive Inference This setting requires that all entities, relation names, and timestamps are known during the training process. The inference graph contains only known entities, relation names, and timestamps (see Fig. 1-a). In this setting, we have $V_{train} = V_{inf}$, $R_{train} = R_{inf}$ and $T_{train} = T_{inf}$. This implies that the entity set, relation name set, and timestamp set remain identical in both the training and test processes.

Semi-Inductive Inference In this setting, the entity and relation name sets are shared between the training and test graphs: $V_{train} = V_{inf}$, $R_{train} = R_{inf}$. Similarly, times are shared between the training and the observed graphs: $T_0 = T_{train}$. However, this task enforces the constraint: $\forall \tau_i \in T_{train}, \forall \tau_j \in T_{valid}, \forall \tau_k \in T_{test}, \tau_i < \tau_j < \tau_k$ (see Fig. 1-b). This ensures that all timestamps in the validation set are strictly later than those in the training set, and all timestamps in the test set are strictly later than

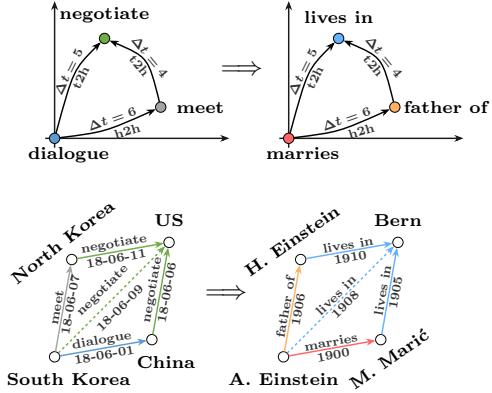


Figure 2: The bottom TKGs depict temporal knowledge graphs from different domains. The upper relation graphs show relative relation representations learned via fundamental interaction patterns (See Section 4.1) and the relative temporal ordering between corresponding facts (Δt) (See Section 4.2) which are transferrable across TKGs. More details are shown in Section E in Appendix.

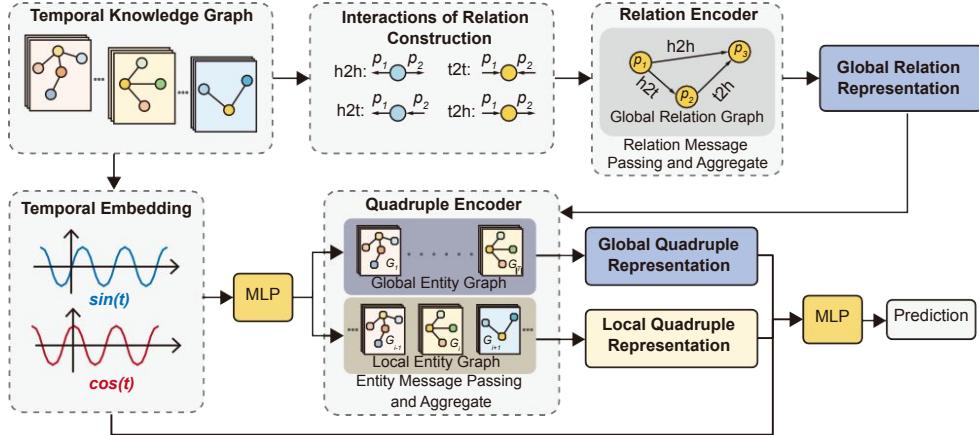


Figure 3: The overall architecture of POSTRA. The model first constructs a Global Relation Graph by treating relations as nodes and the four types of interactions of relations as edges. The Relation Encoder learns global relation representations on the global relation graph through message-passing, which are then used in downstream quadruple encoding (See Section 4.1). Temporal Embedding encodes timestamps’ relative ordering in the TKG by sine and cosine functions (See Section 4.2). The Quadruple Encoder processes entity representations from both Global Entity Graphs and Local Entity Graphs (see Section 4.3), generating Global and Local Quadruple Representations. Finally, these representations, along with temporal embeddings, are fused through an MLP to produce the final Prediction.

those in the validation set, thereby meeting the requirement for predicting future temporal facts. It enables semi-inductive inference on future timestamps.

However, the two existing inference types do not support cross-dataset inductive inference, as the entity and relation sets remain shared across datasets.

Fully-Inductive Inference We propose the more challenging task of fully inductive temporal knowledge graph inference, in which the inference graph contains entities, relation names, and timestamps that never appear in the training data (see Fig. 1-c). This formally means $V_{train} \cap V_{inf} = \emptyset, R_{train} \cap R_{inf} = \emptyset, T_{train} \cap T_{inf} = \emptyset$. Unlike temporal extrapolation inference, this setting does not impose constraints on the sequence of timestamps between the training and inference graphs. In this setting, the pretrained model utilizes weights learned from the training graph to generate adaptive embeddings for the inference graph (with a totally new graph vocabulary). Crucially, a genuine foundation model for temporal knowledge graphs must work in this fully inductive setting and achieve maximal transferability of temporal structural patterns. **To our knowledge, no existing structural temporal knowledge graph learning model works entirely in a fully inductive setting—a gap this paper aims to close as an important step toward foundation models for temporal knowledge graphs.**

3 Related Work

3.1 Temporal Knowledge Graph Embedding

Transductive Models TTransE [Leblay and Chekol, 2018] and TA-DistMult [García-Durán et al., 2018] are among the earliest models to incorporate temporal information into score functions by treating time as an additional element. T(NT)ComplEx [Lacroix et al., 2020] formulates temporal knowledge graph completion as a fourth-order tensor completion problem using semantic matching. Later work TLT-KGE [Zhang et al., 2022] extends it to quaternion space and exchanges information through quaternion operations. HGE [Pan et al., 2024] embeds temporal knowledge graphs into a

product space of suitable manifolds and adopts an attention mechanism that captures the similarities of structural patterns and manifolds. This line of work lacks generalization to unseen TKGs.

Temporal Extrapolation Models Most existing models utilize graph neural network architectures with various updating strategies to predict future facts [Li et al., 2022b,a, Sun et al., 2021, Liang et al., 2023]. For improved explainability, TLogic [Liu et al., 2022] extracts temporal rules using temporal random walks, while xERTE [Han et al., 2020] performs predictions based on explainable subgraphs.

Additionally, TPAR [Chen et al., 2024] is a model capable of handling both transductive and temporal extrapolation tasks. It updates entity representations using Bellman-Ford-based recursive encoding on temporal paths. However, TPAR is unable to perform inductive inference where new entities, relations, and timestamps emerge.

3.2 Inductive Learning on KGs

Structural Inductive Learning Early inductive models, such as NBFNet [Zhu et al., 2021], Grail [Teru et al., 2020], INDIGO [Liu et al., 2021], and Morse [Chen et al., 2022], learn relational patterns in knowledge graphs and can handle unseen entities. However, they require that the relations in the training and inference graphs remain the same. Later works, such as INGRAM [Lee et al., 2023b], TARGI [Ding et al., 2025] and ULTRA [Galkin et al., 2024], construct relation graphs where relations are treated as nodes and interactions between relations as edges. The relations are learned through similar interaction structures utilized for entity representation learning. This enables structural information sharing across training and inference graphs, allowing them to predict unseen entities and relations. However, none of these models can infer temporal facts with unseen timestamps.

Textual Descriptions and LLMs SST-BERT [Chen et al., 2023], ECOLA [Han et al., 2023], and PPT [Xu et al., 2023] utilize BERT-based architectures by pretraining on a TKG corpus, encoding additional textual information of facts, and converting link prediction into masked token prediction, respectively. ICL [Lee et al., 2023a] explores in-context learning of temporal facts within LLMs, while GenTKG [Liao et al., 2024] encodes temporal paths retrieved through logical rule-based methods in LLMs. zrLLM [Ding et al., 2024] first generates relation representations by inputting their textual descriptions into LLMs. All these methods demand substantial computational resources or rely on text descriptions of entities and relations. In contrast, our method is purely structure-driven: it learns directly from the topology of temporal knowledge graphs and does not require any textual annotations for entities or relations.

4 Methodology

To enable fully inductive inference on TKGs, we propose the POSTRA model. The overall architecture of the model is illustrated in Figure 3.

4.1 Relational Structure Learning

To enable the relational transferability across datasets, we adopt the relation encoding method proposed by ULTRA [Galkin et al., 2024]. ULTRA constructs a relation-relation interaction graph for every knowledge graph with four common types of interactions between relations in the knowledge graph: $H = \{h2h, h2t, t2h, t2t\}$, which means head-to-head, head-to-tail, tail-to-head, and tail-to-tail interactions. *Interactions of Relations Construction* part in Fig. 3 illustrates these interactions.

Given a TKG G , we construct a *relation graph* $G_r = (R, H, O)$. In G_r the nodes are the relation names of G , and the edges are the interactions of these relations. For example, $(p_1, h2h, p_2)$ belongs to O if and only if there are two temporal facts of the form (v_1, p_1, v_2, τ_1) and (v_1, p_2, v_3, τ_2) in graph G .

Given a temporal query $(s, p, ?, \tau)$ and a relation graph G_r , we then obtain d -dimensional relation representations $\mathbf{r}_q, q \in R$ via message-passing with neighboring relation nodes:

$$\mathbf{r}_{q|p}^0 = \mathbf{1}_{p=q} * \mathbf{1}^d, \quad q \in R \quad (1)$$

$$\mathbf{r}_{q|p}^{l+1} = \text{AGG} \left(\text{MSG}(\mathbf{r}_{w|p}^l, \mathbf{h}) \mid w \in N_h(p), h \in H \right) \quad (2)$$

where $\mathbf{r}_{q|p}^{l+1}$ stands for the relation q representation at $(l+1)$ -th layer. $N_h(p)$ stands for the neighboring relation nodes of p which are connected by interactions in H . AGG is the aggregation function

(e.g., *sum*), and MSG is the message function (e.g., DistMult’s MSG [Yang et al., 2015] is $\mathbf{r}_{w|p}^l * \mathbf{h}$, TransE’s MSG [Bordes et al., 2013] is $\mathbf{r}_{w|p}^l + \mathbf{h}$), which defines how information is passed from a node’s neighbors to that node. $\mathbf{r}_{q|p}^0$ is initialized by detecting if the relation node is equal to the query relation p . The GNN architecture follows Neural Bellman-Ford Network[Zhu et al., 2021]. As these four interactions are universal and independent of datasets, the relational transferability can be achieved by transferring the embeddings of the four interactions.

4.2 Temporal Embedding

While previous works have used universal structures in graphs as vocabularies for pretraining Graph Foundation Models (GFM) [Sun et al., 2025, Wang et al., 2024], temporal information differs fundamentally from nodes and edges in graphs, as time is inherently sequential. To address this, we represent a temporal knowledge graph (TKG) as a sequence of TKG snapshots G_i , where timestamp index i denotes the position of G_i in the sequence. Given that sinusoidal positional encodings have been shown to effectively capture sequential dependencies, we adopt a similar structure to encode temporal information in TKG sequences. Specifically, we utilize sine and cosine functions with different frequencies to encode the temporal information for G_i with the same dimension size as the proposed model.

$$[\text{TE}(i)]_{2n} = \sin(\omega_n i), \quad [\text{TE}(i)]_{2n+1} = \cos(\omega_n i), \\ \omega_n \in R, \quad 0 \leq n < \frac{d_{\text{PE}}}{2},$$

where n is the dimension index of the embedding. ω_n can be trainable or set to $\omega_n = \beta^{-2n/d_{\text{PE}}}$ where d_{PE} is the dimension size of the temporal embedding and quadruple representation. We also denote representation of τ_i by $\text{TE}_{\tau_i} = \text{TE}(i)$ where i is the timestamp index. The original Transformer architecture sets β as 10000. In this way, each dimension of the positional encoding corresponds to a sinusoidal function, with wavelengths forming a geometric progression from 2π to $10000 \cdot 2\pi$. We chose this function for its computational simplicity and dataset-agnostic design, which preserves temporal transferability. Moreover, prior work has demonstrated that such positional encodings enable the model to effectively learn relative position dependencies [Su et al., 2024, Vaswani et al., 2017], which we argue is crucial for capturing temporal dependencies between temporal facts.

4.3 Temporal-Aware Quadruple Learning

Global Quadruple Representation We incorporate temporal information by integrating the temporal embedding into quadruple-level message passing. For a given temporal query $(s, p, ?, \tau_i)$, we first derive the d -dimensional entity representation $\mathbf{e}_{v|s}, v \in G$ on the entity graph conditioned on the temporal query with head entity s , following a similar approach as described in Section 4.1:

$$\mathbf{e}_{v|s}^0 = \mathbf{1}_{v=s} * \mathbf{r}_p, \quad v \in G, \quad (3)$$

$$\mathbf{e}_{v|s}^{l+1} = \text{AGG} \left(\text{T-MSG}(\mathbf{e}_w^l, \mathbf{r}_q, \sum g^{l+1}(\text{TE}_{\tau_j})) \mid w \in N_q(v), q \in R, \forall \tau_j : (e_w, q, v, \tau_j) \in G \right), \quad (4)$$

where $\mathbf{e}_{v|s}^{l+1}$ is the quadruple representation at the $(l+1)$ th layer, $N_p(s)$ denotes the neighboring entity nodes of s , and TE_{τ_j} embedding of all timestamp indices where (s, p, v, τ_j) occurs in the training graph during training and in the inference graph during testing. $g^{l+1}(\cdot)$ is a linear function defined per layer. T-MSG is a temporal aggregation mechanism that incorporates the temporal embedding described in Section 4.2. In our experiments, we employ TTransE [Leblay and Chekol, 2018], TComplEx [Lacroix et al., 2020], and TNTComplEx [Lacroix et al., 2020] as temporal modeling approaches. The initial entity representation $e_{v|s}^0$ is determined by checking whether the entity node v matches the query entity s . If so, it is initialized using the relation representation \mathbf{r}_p learned in Section 4.1. Since both the relation representations from Section 4.1 and the temporal representations from Section 4.2 are transferable, the learned entity embeddings also satisfy the transferability requirement.

Local Quadruple Representation Relations in TKGs may exhibit different frequencies of change, ranging from fully static to rapidly evolving behaviors [Lacroix et al., 2020]. For example, the relation *CapitalOf* tends to remain stable over time, whereas the relation *Consult* is typically short-term. To effectively model these dynamic relations, which are more influenced by adjacent quadruples,

we construct a local graph for a given temporal query $(s, p, ?, \tau_i) \in G_i$. Specifically, we define the local graph as $G_{local} = \{G_{i-k}, \dots, G_i, \dots, G_{i+k}\}$, where k is time window size. The local quadruple representation $e_{v|s, \tau_i, local}$ is calculated as:

$$\begin{aligned} \mathbf{e}_{v|s, \tau_i, local}^0 &= \mathbf{1}_{v=s} * \mathbf{r}_p, \quad v \in G_{local}, \\ \mathbf{e}_{v|s, \tau_i, local}^{l+1} &= \text{AGG} \left(\text{T-MSG}(\mathbf{e}_{w, local}^l, \mathbf{r}_q, g^{l+1}(TE_{\tau_j}) \mid w \in N_q(v), q \in R, \forall \tau_j \in T (e_w, q, v, \tau_j) \in G_{local}) \right). \end{aligned} \quad (5)$$

Here, we reuse the message-passing structure from Equation 3 to minimize the number of parameters while maintaining consistency in the model architecture.

Prediction Given a temporal query $(s, p, ?, \tau_i)$ and a tail candidate o_{cand} , the final score is:

$$\begin{aligned} V_{s, p, o_{cand}} &= \alpha \mathbf{e}_{o_{cand}|s, \tau_i, local} + (1 - \alpha) \mathbf{e}_{o_{cand}|s}, \\ S(s, p, o_{cand}, \tau_i) &= f_\theta(V_{s, p, o_{cand}}, TE_{\tau_i}), \end{aligned} \quad (6)$$

where $V_{s, p, o_{cand}}$ is the time-aware triple representation (from message passing), $f_\theta(\cdot)$ is a multilayer perceptron with parameters θ , and α is a hyperparameter, providing a tradeoff between local and global information. The method offers theoretical benefits, including the ability to model relations occurring at different temporal frequencies; formal theorems and their proofs are in Appendix I.

5 Experimental Setup

Datasets To evaluate the effectiveness of the proposed model on the fully-inductive inference task, we conduct link prediction experiments on five widely-used temporal knowledge graph benchmark datasets: ICEWS14 [García-Durán et al., 2018], ICEWS05-15 [García-Durán et al., 2018], GDELT [Trivedi et al., 2017], ICEWS18 [Jin et al., 2020], and YAGO [Jin et al., 2020]. These datasets cover a diverse range of domains and temporal characteristics:

1. **Inference Type:** ICEWS14, ICEWS05-15, and GDELT were designed for transductive inference, while ICEWS18 and YAGO were developed for temporal extrapolation tasks.
2. **Domain Coverage:** ICEWS14, ICEWS05-15, and ICEWS18 are subsets of the Integrated Crisis Early Warning System (ICEWS) [Lautenschlager et al., 2015], consisting of news event data. GDELT is a large-scale graph focused on capturing human behavior events. YAGO is derived from the YAGO knowledge base for commonsense facts.
3. **Temporal Granularity:** ICEWS14, ICEWS05-15, and ICEWS18 use daily timestamps. GDELT provides fine-grained 15-minute intervals, while YAGO uses yearly granularity.
4. **Time Span:** ICEWS14 covers the year 2014, ICEWS05-15 spans from 2005 to 2015, and ICEWS18 includes events from January 1 to October 31, 2018. GDELT spans one year from April 2015 to March 2016, while YAGO spans a long range of 189 years.

To assess the model’s fully-inductive inference capabilities, we perform cross-dataset zero-shot evaluations, where the model is trained on one dataset and evaluated on another, ensuring no overlap in entities, relations, or timestamps. Table 6 in the Appendix shows the detailed splits of the datasets.

Baselines We compare POSTRA with two SOTA fully-inductive inference models, INGRAM[Lee et al., 2023b] and ULTRA[Galkin et al., 2024], which can handle unknown entities and relations for fully-inductive inference. We also compare with two LLM-based temporal knowledge graph reasoning models, ICL[Lee et al., 2023a] and GenTKG[Liao et al., 2024], which focus on temporal extrapolation inference task.

Evaluation Metrics We adopt the link prediction task to evaluate our proposed model. Link prediction infers the missing entities for incomplete facts. During the test step, we follow the procedure of [Xu et al., 2020] to generate candidate quadruples. From a test quadruple (s, p, o, τ) , we replace s with all $\bar{s} \in V$ and o with all $\bar{o} \in V$ to get candidate answer quadruples (s, p, \bar{o}, τ) and (\bar{s}, p, o, τ) to queries $(s, p, ?, \tau)$ and $(?, p, o, \tau)$. For each query, all candidate answer quadruples will be ranked by their scores using a time-aware filtering strategy [Goel et al., 2020]. We evaluate our models with four metrics: Mean Reciprocal Rank (MRR), the mean of the reciprocals of predicted ranks of correct

Table 1: Fully-inductive link prediction results. Each block shows training on one dataset and testing on the remaining without fine-tuning. Best results are in **bold**.

Trained on ICEWS14																								
Model	To ICEWS05-15						To GDELT						To ICEWS18						To YAGO				Total Avg	
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10		MRR	H@1	H@10					
INGRAM	6.3	1.3	16.3	7.7	3.0	15.9	10.3	7.0	13.4	17.5	13.2	25.6	10.5	6.1	17.8									
ULTRA	11.1	12.0	15.0	17.0	10.0	30.0	21.8	11.5	43.4	63.7	50.0	91.5	28.4	21.0	45.0									
POSTRA	42.9	35.3	55.3	26.1	17.2	43.8	25.0	14.2	47.3	87.9	83.9	93.2	45.5	37.7	59.9									
Trained on ICEWS05-15																								
Model	To ICEWS14						To GDELT						To ICEWS18						To YAGO				Total Avg	
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10		MRR	H@1	H@10					
INGRAM	11.1	3.8	27.8	3.8	1.0	7.1	6.2	2.0	15.3	12.4	7.6	21.2	8.4	3.6	17.9									
ULTRA	46.5	34.3	70.4	17.9	10.2	32.4	14.5	8.3	36.1	61.2	46.6	90.9	35.0	24.9	57.5									
POSTRA	51.5	41.5	70.6	18.9	10.2	36.7	15.8	8.7	29.7	62.2	49.0	86.5	37.1	27.4	55.9									
Trained on GDELT																								
Model	To ICEWS14						To ICEWS05-15						To ICEWS18						To YAGO				Total Avg	
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10		MRR	H@1	H@10					
INGRAM	8.6	2.9	20.7	5.8	1.2	14.4	5.0	2.3	9.3	7.4	3.3	15.0	6.7	2.4	14.9									
ULTRA	33.6	24.5	50.7	38.3	26.8	60.5	24.4	15.6	41.8	48.5	40.1	63.2	36.2	26.8	54.1									
POSTRA	43.7	32.4	65.7	45.5	33.1	70.0	27.2	16.2	50.4	63.3	50.2	86.8	44.9	33.0	68.0									

quadruples, and Hits@(1/10), the percentage of ranks not higher than 1/10. For all experiments, the higher the better. For ICEWS18 and YAGO, we perform the single-step prediction as mentioned in [Gastinger et al., 2023]. We provide the hyperparameters and training details in Section F in the appendix.

6 Experimental Result

6.1 Fully-Inductive Inference Results

We evaluate POSTRA’s fully-inductive inference performance across multiple zero-shot scenarios, where no fine-tuning is applied during testing, as presented in Table 1. POSTRA consistently outperforms baseline models across multiple evaluation metrics, demonstrating its strong fully-inductive learning capability. From the result, we have following observations:

- 1) POSTRA generalizes well to datasets with varying time spans and temporal granularities. The time span of the experimental datasets ranges from just 1 month (GDELT) to 189 years (YAGO), while their temporal granularities vary from 15 minutes (GDELT) to 1 year (YAGO). Despite being trained on one dataset and evaluated on another with significantly different temporal characteristics, POSTRA consistently achieves strong performance across all scenarios. This demonstrates the robustness and generalization ability of our sequential temporal embedding in handling diverse and unseen temporal information. Furthermore, the performance gap between POSTRA and ULTRA is more pronounced when trained on ICEWS14 compared to ICEWS05-15, suggesting that the proposed sequential temporal embedding is particularly beneficial for smaller datasets.
- 2) POSTRA generalizes well across datasets from different domains and varying densities. The experimental datasets cover a wide range of domains, from encyclopedic knowledge in YAGO to diplomatic event data in ICEWS. POSTRA achieves impressive results even when trained on one domain and evaluated on another (see “Trained on ICEWS14 to YAGO”), demonstrating the model’s ability to transfer across domains. This highlights that the learned quadruple representations from the relation encoder and quadruple encoder are capable of capturing domain-specific structural knowledge. Moreover, the datasets also vary in density: GDELT has more frequent events per timestamp, while ICEWS14 is relatively sparse. POSTRA consistently performs well in cross-dataset evaluation, demonstrating its robustness in event densities.
- 3) POSTRA generalizes well to both temporal knowledge graph interpolation and extrapolation tasks. As shown in Table 2, POSTRA achieves strong results when trained to predict historical unseen events and tested on future unseen events, even outperforming LLM-based models, which are designed for extrapolation tasks. Notably, the cost of LLM-based models to compute the standard MRR metric

over all candidate entities is prohibitive, whereas POSTRA efficiently supports full evaluation. This highlights both the flexibility and the computational efficiency of POSTRA.

Table 2: Link prediction results on ICEWS18 and YAGO between LLM-based models and POSTRA trained on ICEWS14. The best results among all models are in **bold**.

Model	ICEWS18			YAGO		
	MRR	H@1	H@10	MRR	H@1	H@10
ICL	–	18.2	41.4	–	78.4	92.7
GenTKG	–	24.3	42.1	–	79.2	84.3
POSTRA	25.0	14.2	47.3	87.9	83.9	93.2

6.2 Ablation Study

Module Ablation To validate the effectiveness of each module in POSTRA, we conducted ablation studies on three key components: Temporal Embedding (TE), Global Quadruple Representation (GQR), and Local Quadruple Representation (LQR). As shown in Table 3, the results highlight the importance of each component. Removing the Temporal Embedding module results in a substantial performance drop, emphasizing Temporal Embedding’s transferability between different temporal knowledge graphs. Excluding the Global Quadruple Representation significantly reduces the model’s Hits@10 score, supporting our hypothesis that it captures long-term global temporal information. Similarly, removing the Local Quadruple Representation notably decreases the Hits@1 score, aligning with our assumption that it models the concurrent interactions of temporal events. We provide a more detailed visualization of GQR and LQR in Section H in the Appendix.

Performance on Temporal Structural Patterns To further assess the model’s ability to capture temporal structural patterns, we evaluate on two temporal structural patterns: symmetric and inverse. We construct subsets from the ICEWS05-15 test set that contain such structural patterns¹ and report results in Table 4. A pair of quadruples is *symmetric* if for (s, p, o, τ_1) , (o, p, s, τ_2) exists. A relation pair (p, p') is *inverse* if (s, p, o, τ_1) implies (o, p', s, τ_2) . For both patterns, POSTRA consistently outperforms the baseline ULTRA. These results demonstrate that incorporating relative temporal signals significantly enhances the model’s ability to learn and generalize structural patterns in TKGs.

Table 4: Model performance on structural pattern subsets. All models are trained on ICEWS14.

Model	symmetric			inverse		
	MRR	H@1	H@10	MRR	H@1	H@10
ULTRA	53.5	38.8	81.8	45.7	32.5	70.9
POSTRA	67.0	53.7	90.7	61.8	48.9	86.0

7 Conclusion

We introduce POSTRA, which advances beyond existing transductive or semi-inductive approaches to support fully inductive inference. POSTRA exhibits strong universal transferability across diverse time spans, temporal granularities, domains, and prediction tasks, overcoming limitations of prior models that depended on dataset-specific entity, relation, and timestamp embeddings. Extensive evaluations across multiple scenarios show that POSTRA consistently outperforms existing state-

¹Symmetric and inverse subsets contain 12,600 and 28,891 quadruples, respectively. Data splits are provided in our codebase under dataset classes ICEWS0515_sym and ICEWS0515_inv.

of-the-art models in zero-shot settings. In future work, we aim to extend POSTRA to support time prediction tasks and explore richer temporal representations.

Acknowledgment

This work has received funding from the CHIPS Joint Undertaking (JU) under grant agreement No. 101140087 (SMARTY). The JU receives support from the European Union’s Horizon Europe research and innovation programme. Furthermore, on national level this work is supported by the German Federal Ministry of Education and Research (BMBF) under the sub-project with the funding number 16MEE0444. The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (<https://www.nhr-verein.de/en/our-partners>). HoreKa is partly funded by the German Research Foundation (DFG).

References

- Jing Ao, Jon Doyle, Christopher Healey, and Ranga Raju Vatsavai. *Temporal Knowledge Graphs: Integration, Querying, and Analytics*. PhD thesis, North Carolina State University, 2022. AAI30283568.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.
- Li Cai, Xin Mao, Yuhao Zhou, Zhaoguang Long, Changxu Wu, and Man Lan. A survey on temporal knowledge graph: Representation learning and applications, 2024. URL <https://arxiv.org/abs/2403.04782>.
- Kai Chen, Ye Wang, Yitong Li, Aiping Li, Han Yu, and Xin Song. A unified temporal knowledge graph reasoning model towards interpolation and extrapolation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 117–132, 2024.
- Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. Meta-knowledge transfer for inductive knowledge graph embedding. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 927–937, 2022.
- Zhongwu Chen, Chengjin Xu, Fenglong Su, Zhen Huang, and Yong Dou. Incorporating structured sentences with time-enhanced bert for fully-inductive temporal relation prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 889–899, 2023.
- Ling Ding, Lei Huang, Zhizhi Yu, Di Jin, and Dongxiao He. Towards global-topology relation graph for inductive knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11581–11589, 2025.
- Zifeng Ding, Heling Cai, Jingpei Wu, Yunpu Ma, Ruotong Liao, Bo Xiong, and Volker Tresp. zrlm: Zero-shot relational learning on temporal knowledge graphs with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1877–1895, 2024.
- Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. Towards foundation models for knowledge graph reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.
- Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4816–4821, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1516. URL <https://aclanthology.org/D18-1516/>.
- Julia Gastinger, Timo Sztyler, Lokesh Sharma, Anett Schuelke, and Heiner Stuckenschmidt. Comparing apples and oranges? on the evaluation of methods for temporal knowledge graph forecasting. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 533–549. Springer, 2023.
- Rishab Goel, Seyed Mehran Kazemi, Marcus Brubaker, and Pascal Poupart. Diachronic embedding for temporal knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3988–3995, 2020.
- Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *International conference on learning representations*, 2020.

- Zhen Han, Ruotong Liao, Jindong Gu, Yao Zhang, Zifeng Ding, Yujia Gu, Heinz Koepll, Hinrich Schütze, and Volker Tresp. Ecola: Enhancing temporal knowledge embeddings with contextualized language representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5433–5447, 2023.
- Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inference over temporal knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6669–6683, 2020.
- Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. Tensor decompositions for temporal knowledge base completion. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rke2P1BFwS>.
- Jennifer Lautenschlager, Steve Shellman, and Michael Ward. Icews event aggregations. *Harvard Dataverse*, 3(595):28, 2015.
- 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.
- Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. Temporal knowledge graph forecasting without knowledge using in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 544–557, 2023a.
- Jaejun Lee, Chanyoung Chung, and Joyce Jiyoung Whang. Ingram: Inductive knowledge graph embedding via relation graphs. In *International Conference on Machine Learning*, pages 18796–18809. PMLR, 2023b.
- 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*, pages 2152–2158, 2022a.
- Zixuan Li, Saiping Guan, Xiaolong Jin, Weihua Peng, Yajuan Lyu, Yong Zhu, Long Bai, Wei Li, Jiafeng Guo, and Xueqi Cheng. Complex evolutional pattern learning for temporal knowledge graph reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 290–296, 2022b.
- 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 *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1559–1568, 2023.
- Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. Gentkg: Generative forecasting on temporal knowledge graph with large language models. In *NAACL-HLT (Findings)*, 2024.
- Shuwen Liu, Bernardo Grau, Ian Horrocks, and Egor Kostylev. Indigo: Gnn-based inductive knowledge graph completion using pair-wise encoding. *Advances in Neural Information Processing Systems*, 34:2034–2045, 2021.
- Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4120–4127, 2022.
- Jixin Pan, Mojtaba Nayyeri, Yinan Li, and Steffen Staab. Hge: embedding temporal knowledge graphs in a product space of heterogeneous geometric subspaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8913–8920, 2024.
- Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6663–6676, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.520. URL <https://aclanthology.org/2021.acl-long.520/>.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

- Haohai Sun, Jialun Zhong, Yunpu Ma, Zhen Han, and Kun He. Timetraveler: Reinforcement learning for temporal knowledge graph forecasting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8306–8319, 2021.
- Li Sun, Zhenhao Huang, Suyang Zhou, Qiqi Wan, Hao Peng, and Philip S Yu. Riemanngfm: Learning a graph foundation model from structural geometry. In *THE WEB CONFERENCE 2025*, 2025.
- Komal Teru, Etienne Denis, and Will Hamilton. Inductive relation prediction by subgraph reasoning. In *International conference on machine learning*, pages 9448–9457. PMLR, 2020.
- Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *international conference on machine learning*, pages 3462–3471. PMLR, 2017.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Zehong Wang, Zheyuan Zhang, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. Gft: Graph foundation model with transferable tree vocabulary. *Advances in Neural Information Processing Systems*, 37: 107403–107443, 2024.
- Chengjin Xu, Mojtaba Nayyeri, Fouad Alkhoury, Hamed Shariat Yazdi, and Jens Lehmann. Tero: A time-aware knowledge graph embedding via temporal rotation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1583–1593, 2020.
- Wenjie Xu, Ben Liu, Miao Peng, Xu Jia, and Min Peng. Pre-trained language model with prompts for temporal knowledge graph completion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7790–7803, 2023.
- Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*, 2015.
- Fuwei Zhang, Zhao Zhang, Xiang Ao, Fuzhen Zhuang, Yongjun Xu, and Qing He. Along the time: Timeline-traced embedding for temporal knowledge graph completion. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2529–2538, 2022.
- Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural bellman-ford networks: A general graph neural network framework for link prediction. *Advances in neural information processing systems*, 34:29476–29490, 2021.

A Parameter Count and Complexity

From Table 5, we observe that the number of parameters in POSTRA is agnostic to the dataset size. Since POSTRA does not initialize embeddings based on $|V|$, $|R|$, or $|T|$, but instead relies on a fixed number of relation interactions $|H|$, its parameter count is only related to the embedding dimension d . This property makes POSTRA especially suitable for transfer learning scenarios.

The time complexity of POSTRA is primarily determined by the quadruple encoder, as the number of relations $|R|$ is significantly smaller than the number of entities $|V|$, allowing us to omit the complexity contribution of the relation encoder. Furthermore, since the local entity graph is much smaller than the global entity graph, the computational upper bound is dominated by the global quadruple representation. Utilizing NBFNet [Zhu et al., 2021] as the encoder, the time complexity for a single layer is $O(|Q|d + |V|d^2)$. For A layers, the total time complexity becomes $O(A(|Q|d + |V|d^2))$.

The memory complexity of POSTRA is also linear in the number of edges, expressed as $O(A|Q|d)$, as the quadruple encoder maintains and updates representations for each edge in the temporal knowledge graph.

Table 5: Parameter number and average training time for POSTRA.

Model	Datasets	d	Parameter number	Batch Size	Average epoch time(h)
POSTRA	ICEWS14	64	247,937	16	1.8
	ICEWS05-15	64	247,937	2	21.5
	GDELT	64	247,937	1	123.3

Table 6: Detailed statistics of datasets.

Dataset	ICEWS14	ICEWS05-15	GDELT	ICEWS18	YAGO
Entities	7,128	10,488	500	23,033	10,623
Relations	230	251	20	256	10
Times	365	4017	366	303	189
Train	72,826	386,962	2,735,685	373,018	161,540
Validation	8,941	46,275	341,961	45,995	19,523
Test	8,963	46,092	341,961	49,995	20,026
Granularity	Daily	Daily	15 minutes	Daily	Yearly

B Pre-training Results

Table 7 presents the pretrained performance of POSTRA. It achieves comparable performance with TComplEx while using significantly fewer parameters especially on datasets which involve a large number of entities and relations. POSTRA generates adaptive representations dynamically through a conditional message passing mechanism. This design enables effective generalization with a compact model size, making it highly scalable and suitable for large temporal knowledge graphs.

Table 7: Pretrained results of POSTRA.

Model	ICEWS14			ICEWS-0515			GDELT		
	MRR	H@10	Parameter Count	MRR	H@10	Parameter Count	MRR	H@10	Parameter Count
TComplEx	56.2	73.2	2,035,968	58.0	78.1	3,841,792	21.4	40.1	251,936
POSTRA	58.3	76.0	247,937	59.0	78.7	274,937	24.5	42.7	247,937

C Limitations

Although POSTRA demonstrates strong fully-inductive inference performance in zero-shot scenarios, several limitations remain. First, training on larger graphs does not always lead to improved performance. We hypothesize that variations in temporal patterns across datasets of different sizes may influence results. Second, POSTRA can be computationally intensive for dense temporal knowledge graphs, as its space complexity scales with the number of events. Third, POSTRA is unable to perform time prediction. We aim to explore more efficient architectures and adaptive training strategies to address these challenges in future work.

D Ethics Statement

POSTRA can be applied to a wide range of tasks and datasets beyond their originally intended scope. Given the sensitive and time-dependent nature of temporal knowledge graphs (TKGs), there is a risk that such models could be used to infer patterns from historical or anonymized temporal data in ways that may compromise privacy or be exploited for surveillance, misinformation, or manipulation. On the positive side, fully-inductive models for TKGs offer significant benefits by enabling zero-shot transfer across domains, datasets, and temporal granularities. This reduces the need to train separate models for each new scenario, lowering the computational cost for training repeated models.

E Relative Representation Transfer

Subgraph (a) in Figure 4 shows the transfer of relative entity representations. In the left bottom KG, $(\text{South Korea} \xrightarrow{\text{meet}} \text{North Korea} \xrightarrow{\text{negotiate}} \text{US}) \wedge (\text{South Korea} \xrightarrow{\text{dialogue with}} \text{China} \xrightarrow{\text{negotiate}} \text{US})$ implies that $\text{South Korea} \xrightarrow{\text{negotiate}} \text{US}$. Given entity v_1, v_2, v_3, v_4 , we can learn a structure that $(v_1 \xrightarrow{\text{meet}} v_2 \xrightarrow{\text{negotiate}} v_3) \wedge (v_1 \xrightarrow{\text{dialogue with}} v_4 \xrightarrow{\text{negotiate}} v_3)$ implies that $v_1 \xrightarrow{\text{negotiate}} v_3$ during training. In the right part, we substitute v_1, v_2, v_3, v_4 with Germany, India, UK and EU respectively, then we can infer that $\text{Germany} \xrightarrow{\text{negotiate}} \text{UK}$.

Subgraph (b) depicts the transfer of relative relation representations. In the left part, the upper relation graph depicts the interaction of relations in the bottom KG: $(\text{dialogue with} \xrightarrow{\text{h2h}} \text{meet} \xrightarrow{\text{t2h}} \text{negotiate}) \wedge (\text{dialogue with} \xrightarrow{\text{t2h}} \text{negotiate})$. Given relation r_1, r_2, r_3 , we learn a relational structure: $(r_1 \xrightarrow{\text{h2h}} r_2 \xrightarrow{\text{t2h}} r_3) \wedge (r_1 \xrightarrow{\text{t2h}} r_3)$ during training. In the right part, we substitute r_1, r_2, r_3 with *marries*, *father of* and *lives in* respectively, then we can the relational structure for the KG in the right part.

Subgraph (c) depicts the transfer of relative temporal ordering. In the left part, the upper relation graph depicts the relative temporal ordering of quadruples in the bottom TKG: $(\text{dialogue with} \xrightarrow{\Delta=6} \text{meet} \xrightarrow{\Delta=4} \text{negotiate}) \wedge (\text{dialogue with} \xrightarrow{\Delta=5} \text{negotiate})$. Given relation r_1, r_2, r_3 , we learn a relational structure: $(r_1 \xrightarrow{\Delta=6} r_2 \xrightarrow{\Delta=4} r_3) \wedge (r_1 \xrightarrow{\Delta=5} r_3)$ during training. In the right part, we substitute r_1, r_2, r_3 with *marries*, *father of* and *lives in* respectively, then we can the temporal transition information for the TKG in the right part.

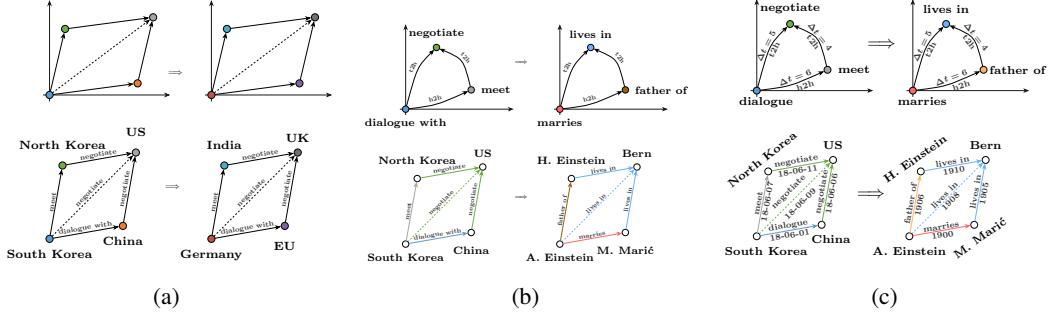


Figure 4: Relative Representation Transfer

F Hyperparameter and Training Details

Our experiments were conducted on 4 NVIDIA A100 GPUs with 48GB of RAM. We set the maximal training epoch as 10 and negative samples as 512. We use a batch size of 16, 2, and 1 for training ICEWS14, ICEWS05-15 and GDELT respectively. We set the dimension size d as 64. We employed the AdamW optimizer and set the learning rate as 0.0005. More details can be seen in Table 8. For the baseline methods, we utilized the settings in their original papers.

Table 8: Detailed hyperparameters. GNN_r denotes the relation representation encoder, GNN_q is quadruple representation encoder.

Module	Hyperparameter	Pre-training
GNN_r	# layers	6
	hidden dim	64
	MSG	Dual
	AGG	sum
GNN_e	# layers	6
	hidden dim	64
	T-MSG	TComplEx
	AGG	sum
k	k	ICEWS14:0; ICEWS0515, GDELT:1
	β	10,000
Prediction	f_θ	2-layer MLP
	α	ICEWS14:0.5; ICEWS0515, GDELT:0.8

G Hyperparameter Analysis

We conduct a sensitivity analysis of three key hyperparameters, k , α , and β on model performance by pre-training on the ICEWS14 dataset. Figure 5 presents the MRR and H@10 results under different values of each hyperparameter while keeping all others fixed at their optimal settings.

Effect of k . As shown in Figure 5(a), increasing k , which controls the length of the local entity graph, leads to a gradual decline in both MRR and H@10. As ICEWS14 is a news dataset and contains mainly short-term temporal relations, deeper propagation may cause oversmoothing and decrease the model’s performance.

Effect of α . Figure 5(b) shows that performance is highly sensitive to the choice of α , which balances the contribution of local and global temporal contexts. Both MRR and H@10 significantly improve as α increases from 0 to 0.5, indicating the benefit of incorporating both types of temporal information. However, performance degrades when α approaches 1, implying that overemphasis on either context harms generalization. The results highlight the importance of maintaining a moderate balance between local and global temporal patterns.

Effect of β . As illustrated in Figure 5(c), the model is highly robust to the choice of β , which controls the sinusoidal frequency of temporal embeddings. Across a wide range of values—from 10^2 to 10^6 —MRR and H@10 remain largely stable, with negligible performance fluctuations.

The hyperparameter analysis reveals that α is the most critical parameter and requires careful tuning to balance local and global temporal information. The choice of k should be dataset-dependent; for datasets characterized by short-term temporal patterns, smaller values of k are preferable. In contrast, the model is relative insensitive to β , so we set it as 10,000 across all datasets.

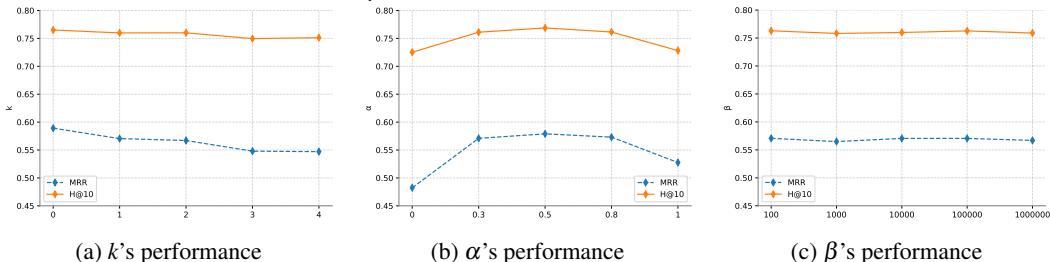


Figure 5: MRR and H@10 performance with different k , α and β . We report ICEWS14’s pre-training performance and keep all other hyper-parameters as the fixed best setting.

H Case Study

To intuitively demonstrate POSTRA’s capability in generating compact and distinguishable embeddings, we employ PCA to visualize the distribution of embeddings for five representative quadruples:

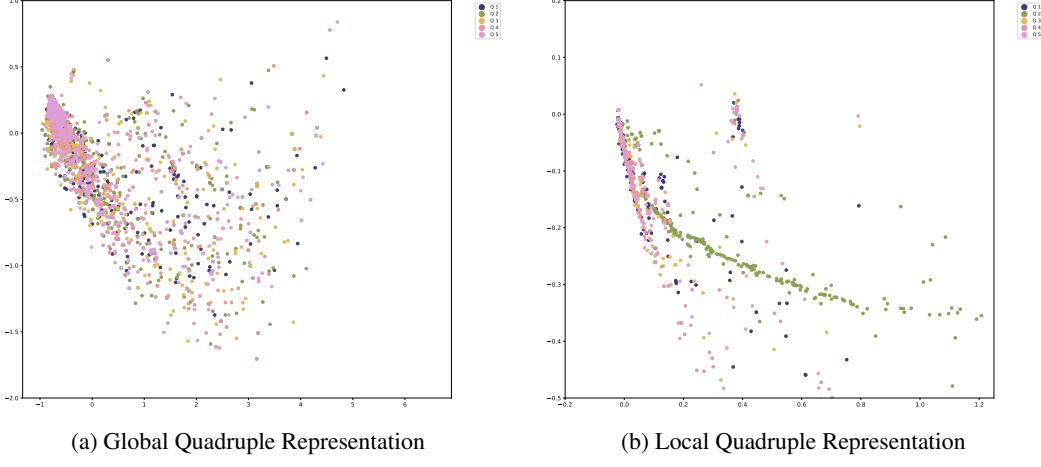


Figure 6: PCA visualizations of five quadruples with (China, Host a visit, ?, ?)

Q1: (China, Host a Visit, Macky Sall, 2014-02-19), Q2:(China, Host a Visit, Teo Chee Hean, 2014-10-28), Q3: (China, Host a Visit, Chuck Hagel, 2014-04-08), Q4: (China, Host a Visit, Sar Kheng, 2014-06-03), and Q5: (China, Host a Visit, Barack Obama, 2014-12-25), as illustrated in Figure 6. The visualization reveals that the Global Quadruple Representation tends to capture general relational patterns, producing embeddings that are closely clustered for the same query structure (China, Host a Visit, ?, ?). In contrast, the Local Quadruple Representation is more sensitive to immediate temporal context, resulting in more distinctive embeddings across different time points. This dual representation mechanism highlights POSTRA’s strength in capturing both global structural semantics and local temporal dynamics. For example, for Q5, Barack Obama is frequently the tail entity for similar queries, leading the global encoder to assign him a high probability and correctly predict the answer. Meanwhile, in Q1, a supporting local event (Macky Sall, Make a Visit, China, 2014-02-18) appears in the surrounding context, enabling the local quadruple encoder to identify Macky Sall correctly. However, the absence of any relevant local events involving Teo Chee Hean in the Q2 causes the model to mispredict Barack Obama.

I Theoretical Analysis

In this section, we provide a theoretical analysis demonstrating the model’s ability to transfer across time and effectively capture a variety of periodic patterns.

Temporal Transferability As illustrated in Figure 2 and 4, the temporal difference is the transferable information in TKGs. Therefore, we expect that such transferability must be followed in the time embedding space. Here, we aim to show that if four time points’ indices— τ_1, τ_2 in the training graph and τ'_1, τ'_2 in the test graph—hold $\Delta T = \tau_2 - \tau_1 = \Delta T' = \tau'_2 - \tau'_1$, then we expect that such transferability is preserved via sinusoidal positional encoding in the form of Euclidean distance $\|\text{TE}(\tau_2) - \text{TE}(\tau_1)\| = \|\text{TE}(\tau'_2) - \text{TE}(\tau'_1)\|$ that is governed solely by their time difference ΔT , rather than the absolute times.

The following theorem states and proves this:

Theorem 1 (Time-Shift Invariance in Sinusoidal Positional Embeddings). *Let $d \in \mathbb{N}$ be even and fix positive frequencies $\omega_0, \omega_1, \dots, \omega_{\frac{d}{2}-1} > 0$. Define the sinusoidal temporal embedding $\text{TE} : \mathbb{N} \rightarrow \mathbb{R}^d$ component-wise by*

$$[\text{TE}(t)]_{2n} = \sin(\omega_n t), \quad [\text{TE}(t)]_{2n+1} = \cos(\omega_n t), \quad 0 \leq n < \frac{d}{2}.$$

For any four time-points indices $\tau_1, \tau_2, \tau'_1, \tau'_2 \in \mathbb{N}$ set

$$\Delta_\tau := \tau_2 - \tau_1, \quad \Delta'_\tau := \tau'_2 - \tau'_1.$$

If $\Delta_\tau = \Delta'_\tau$ then

$$\|\text{TE}(\tau_2) - \text{TE}(\tau_1)\| = \|\text{TE}(\tau'_2) - \text{TE}(\tau'_1)\|.$$

That is, the Euclidean distance between two sinusoidal embeddings depends only on the time difference, not on the absolute timestamps.

Step-by-step proof.

Fix an index $n \in \{0, \dots, \frac{d}{2} - 1\}$ and two times $t_1, t_2 \in \mathbb{N}$. Denote

$$\Delta = t_2 - t_1, \quad \Sigma = t_1 + t_2.$$

Using the standard sum-and-difference identities,

$$\sin(\omega_n t_2) - \sin(\omega_n t_1) = 2 \sin\left(\frac{\omega_n \Delta}{2}\right) \cos\left(\frac{\omega_n \Sigma}{2}\right),$$

$$\cos(\omega_n t_2) - \cos(\omega_n t_1) = -2 \sin\left(\frac{\omega_n \Delta}{2}\right) \sin\left(\frac{\omega_n \Sigma}{2}\right).$$

Thus, the squared contribution of this frequency to the Euclidean distance is

$$[\sin(\omega_n t_2) - \sin(\omega_n t_1)]^2 + [\cos(\omega_n t_2) - \cos(\omega_n t_1)]^2 = 4 \sin^2\left(\frac{\omega_n \Delta}{2}\right),$$

because $\cos^2 + \sin^2 = 1$ removes any dependence on Σ .

Summing over all n gives

$$\|\text{TE}(t_2) - \text{TE}(t_1)\|^2 = \sum_{n=0}^{\frac{d}{2}-1} 4 \sin^2\left(\frac{\omega_n \Delta}{2}\right) = 4 \sum_{n=0}^{\frac{d}{2}-1} \sin^2\left(\frac{\omega_n \Delta}{2}\right),$$

which is a function of $\Delta = t_2 - t_1$ alone. Consequently, if $\tau_2 - \tau_1 = \tau'_2 - \tau'_1$ then $\Delta_\tau = \Delta'_{\tau'}$ and

$$\|\text{TE}(\tau_2) - \text{TE}(\tau_1)\|^2 = 4 \sum_n \sin^2\left(\frac{\omega_n \Delta_\tau}{2}\right) = 4 \sum_n \sin^2\left(\frac{\omega_n \Delta'_{\tau'}}{2}\right) = \|\text{TE}(\tau'_2) - \text{TE}(\tau'_1)\|^2.$$

Taking square roots results in the claimed equality of norms. \square

Capturing Diverse Periodicities Temporal facts expressed as (s, p, o, τ) , where $\tau \in \mathbb{T}$ represents the time annotation, can capture periodic phenomena, such as the Olympic Games occurring every four years or annual events like the Nobel Prize ceremonies. Modeling diverse periodic behaviors is crucial in the zero-shot transfer learning, as it enables systems to accurately represent and analyze the dynamic evolution of diverse relationships and events. In this section, we prove that POSTRA is capable of capturing periodic (with different frequencies) temporal facts. In the following theorem, we prove that the scorer in Equation 6 can capture and express various frequencies even if we assume there f_θ is simply a linear function (with m linear nodes for representing m different frequencies). For simplicity, we loosely use τ as the timestamp index i in τ_i in the following.

Theorem 2 (Multi-Frequency Affine Scorer). *Let*

- $T > 1$ be an integer time horizon ;
- $d_{\text{PE}} \geq T$ be an even positional-encoding dimension;
- P_1, \dots, P_m , such that $P_i \mid T$ for every $i = 1, \dots, m$, is a family of positive integers.
- $[\text{TE}(\tau)]_{2n} = \sin(\omega_n \tau), \quad [\text{TE}(\tau)]_{2n+1} = \cos(\omega_n \tau), \quad 0 \leq n < \frac{d_{\text{PE}}}{2}$, with arbitrary real frequencies $\omega_0, \dots, \omega_{d_{\text{PE}}/2-1}$, be (standard) sinusoidal positional encoding defined for any τ .
- $G = \{V_1, \dots, V_m\} \subset \mathbb{R}^{dv}$ be m fixed context vectors and,
- $g_i : \{0, 1, \dots, T-1\} \rightarrow \mathbb{R}$ be a non-constant sequence for each $i = 1, \dots, m$, that we extend to all \mathbb{Z} by P_i -periodicity:

$$g_i(\tau) := g_i(\tau \bmod P_i), \quad \tau \in \mathbb{Z}.$$

• and

$$M = \begin{pmatrix} \text{TE}(0)^\top \\ \text{TE}(1)^\top \\ \vdots \\ \text{TE}(T-1)^\top \end{pmatrix} \in \mathbb{R}^{T \times d_{\text{PE}}}, \quad B^{(i)} = \begin{pmatrix} \text{TE}(P_i)^\top - \text{TE}(0)^\top \\ \text{TE}(P_i+1)^\top - \text{TE}(1)^\top \\ \vdots \\ \text{TE}(T-1+P_i)^\top - \text{TE}(T-1)^\top \end{pmatrix} \in \mathbb{R}^{T \times d_{\text{PE}}}.$$

be encoding matrices, we stack $B^{(i)}$ block-diagonally:

$$\tilde{B} := \text{diag}(B^{(1)}, \dots, B^{(m)}) \in \mathbb{R}^{mT \times md_{\text{PE}}}.$$

Define

$$g := \begin{pmatrix} g_1 \\ \vdots \\ g_m \end{pmatrix} \in \mathbb{R}^{mT}, \quad w_{\text{PE}} := \begin{pmatrix} w_{\text{PE}}^{(1)} \\ \vdots \\ w_{\text{PE}}^{(m)} \end{pmatrix} \in \mathbb{R}^{md_{\text{PE}}}.$$

(A) **Compatibility** (\mathcal{C}). There exists $w_{\text{PE}} \in \mathbb{R}^{md_{\text{PE}}}$ solving the linear system

$$\boxed{\begin{pmatrix} I_m \otimes M \\ \tilde{B} \end{pmatrix} w_{\text{PE}} = \begin{pmatrix} g \\ 0 \end{pmatrix}} \quad (\mathcal{C})$$

where $I_m \otimes M$ is the block-diagonal Kronecker product whose i -th diagonal block equals M .

(B) **Existence of a scorer** (\mathcal{S}). There exist parameters

$$\theta = (W_{\text{PE}}, w_V, b) \in \mathbb{R}^{d_{\text{PE}} \times m} \times \mathbb{R}^{d_V} \times \mathbb{R}$$

such that the affine map

$$f_\theta(V, \text{TE}(\tau)) := w_V^\top V + W_{\text{PE}}^\top \text{TE}(\tau) + b \in \mathbb{R}^m$$

obeys, for each $i = 1, \dots, m$,

- (i) P_i -periodicity in $\tau \in \mathbb{Z}$;
- (ii) non-constancy as a function of τ ;
- (iii) interpolation on the basic window:

$$[f_\theta(V_i, \text{TE}(\tau))]_i = w_V^\top V_i + g_i(\tau), \quad 0 \leq \tau < T.$$

Given \mathcal{C} and \mathcal{S} , the following holds

$$\boxed{\mathcal{C} \iff \mathcal{S}}.$$

When these equivalent conditions hold one may choose any solution w_{PE} of (\mathcal{C}) , set

$$W_{\text{PE}} := [w_{\text{PE}}^{(1)} \cdots w_{\text{PE}}^{(m)}], \quad b = 0,$$

and pick an arbitrary w_V ; the resulting θ satisfies \mathcal{S} .

Step-by-step proof of $\mathcal{C} \iff \mathcal{S}$. We write $w_{\text{PE}} = (w_{\text{PE}}^{(1)}, \dots, w_{\text{PE}}^{(m)})$ with each block $w_{\text{PE}}^{(i)} \in \mathbb{R}^{d_{\text{PE}}}$.

Part I: $\mathcal{C} \implies \mathcal{S}$. Assume a vector w_{PE} satisfies system (\mathcal{C}) . Here, we show that if the compatibility condition holds, then a scorer exists. We divide the proof into four parts as follows:

Step 1: The top block $(I_m \otimes M)w_{\text{PE}} = g$ is equivalent to

$$M w_{\text{PE}}^{(i)} = g_i \Big|_{0 \leq \tau < T}, \quad i = 1, \dots, m. \quad (1)$$

Step 2: The block $B w_{\text{PE}} = 0$ gives for every i

$$B^{(i)} w_{\text{PE}}^{(i)} = 0 \implies w_{\text{PE}}^{(i)\top} (\text{TE}(\tau + P_i) - \text{TE}(\tau)) = 0, \quad 0 \leq \tau < T.$$

Thus

$$w_{\text{PE}}^{(i)\top} \text{TE}(\tau + P_i) = w_{\text{PE}}^{(i)\top} \text{TE}(\tau), \quad 0 \leq \tau < T. \quad (2)$$

Replacing τ by $\tau + P_i$ and iterating k times shows

$$w_{\text{PE}}^{(i)\top} \text{TE}(\tau + kP_i) = w_{\text{PE}}^{(i)\top} \text{TE}(\tau), \quad \forall \tau \in \{0, \dots, T-1\}, k \in \mathbb{Z},$$

i.e., P_i -periodicity on all integers.

Step 3: Combine (1) with periodicity: for every $\tau \in \mathbb{Z}$

$$w_{\text{PE}}^{(i)\top} \text{TE}(\tau) = g_i(\tau \bmod P_i). \quad (3)$$

Step 4: Define

$$W_{\text{PE}} := [w_{\text{PE}}^{(1)} \cdots w_{\text{PE}}^{(m)}], \quad b := 0, \quad \text{choose any } w_V \in \mathbb{R}^{d_V}.$$

Then

$$f_\theta(V, \text{TE}(\tau)) = w_V^\top V + \begin{pmatrix} w_{\text{PE}}^{(1)\top} \text{TE}(\tau) \\ \vdots \\ w_{\text{PE}}^{(m)\top} \text{TE}(\tau) \end{pmatrix},$$

whose i -th component equals the right-hand side of (3). Therefore:

- a) it is P_i -periodic by construction; b) it is non-constant because each g_i is non-constant; c) on $0 \leq \tau < T$, (3) reduces to $w_V^\top V_i + g_i(\tau)$.

Hence \mathcal{S} holds.

Part II: $\mathcal{S} \implies \mathcal{C}$.

Conversely, suppose parameters $\theta = (W_{\text{PE}}, w_V, b)$ satisfy \mathcal{S} . Write the columns as $W_{\text{PE}} = [w_{\text{PE}}^{(1)} | \cdots | w_{\text{PE}}^{(m)}]$.

Step 1: Evaluating property (iii) at $0 \leq \tau < T$ gives

$$M w_{\text{PE}}^{(i)} = g_i \Big|_{0 \leq \tau < T}, \quad i = 1, \dots, m. \quad (4)$$

Step 2: P_i -periodicity implies the $B^{(i)}$ equations. Property (i) yields, for every integer τ , $w_{\text{PE}}^{(i)\top} \text{TE}(\tau + P_i) = w_{\text{PE}}^{(i)\top} \text{TE}(\tau)$. Specialising to $\tau = 0, \dots, T-1$ one obtains $B^{(i)} w_{\text{PE}}^{(i)} = 0$.

Step 3: Collecting (4) for all i and the $B^{(i)}$ equations in a single vector w_{PE} yields exactly system (\mathcal{C}) . Thus \mathcal{C} holds.

Both directions are proven; therefore $\mathcal{C} \iff \mathcal{S}$. The final comment about choosing (W_{PE}, w_V, b) is precisely the construction in Part I, Step 4. \square

Following Theorem 2, we are interested in knowing under which conditions the compatibility condition holds itself, i.e., when the system has a solution.

Theorem 3 (Universal Compatibility under harmonically-aligned frequencies). *Fix*

$$T > 1, \quad d_{\text{PE}} \geq T, \quad P_1, \dots, P_m \in \mathbb{N}, \quad P_i \mid T.$$

Let $L := \text{lcm}(P_1, \dots, P_m)^2$ and choose any collection of integers $k_0, k_1, \dots, k_{d_{\text{PE}}/2-1}$. Define the frequencies

$$\omega_n := \frac{2\pi k_n}{L}, \quad 0 \leq n < \frac{d_{\text{PE}}}{2},$$

and build the sinusoidal positional encoding

$$[\text{TE}(\tau)]_{2n} = \sin(\omega_n \tau), \quad [\text{TE}(\tau)]_{2n+1} = \cos(\omega_n \tau), \quad \tau \in \mathbb{N}.$$

Form the matrices as follows

$$M = \begin{pmatrix} \text{TE}(0)^\top \\ \vdots \\ \text{TE}(T-1)^\top \end{pmatrix} \in \mathbb{R}^{T \times d_{\text{PE}}}, \quad B^{(i)} = \begin{pmatrix} \text{TE}(P_i)^\top - \text{TE}(0)^\top \\ \vdots \\ \text{TE}(T-1+P_i)^\top - \text{TE}(T-1)^\top \end{pmatrix} \in \mathbb{R}^{T \times d_{\text{PE}}}.$$

For these frequencies, every $B^{(i)}$ vanishes identically; hence, the stacked compatibility system

$$\begin{pmatrix} I_m \otimes M \\ \text{diag}(B^{(1)}, \dots, B^{(m)}) \end{pmatrix} w_{\text{PE}} = \begin{pmatrix} g \\ 0 \end{pmatrix} \quad (\mathcal{C})$$

is always solvable, regardless of the choice of any target sequences $g_i: \{0, \dots, T-1\} \rightarrow \mathbb{R}$.

We present the Step-by-step proof as follows. **Step 1. Encoding is L -periodic.** Because $\omega_n L = 2\pi k_n$,

$$\sin(\omega_n(\tau+L)) = \sin(\omega_n \tau), \quad \cos(\omega_n(\tau+L)) = \cos(\omega_n \tau), \quad \forall \tau \in \mathbb{Z},$$

so $\text{TE}(\tau+L) = \text{TE}(\tau)$.

Step 2. Encoding is also P_i -periodic. Each P_i divides L , hence $\text{TE}(\tau+P_i) = \text{TE}(\tau)$. Therefore

$$B^{(i)} = 0 \in \mathbb{R}^{T \times d_{\text{PE}}}, \quad i = 1, \dots, m.$$

Step 3. Compatibility degenerates to a single block. With every $B^{(i)}$ equal to zero, system (\mathcal{C}) becomes

$$(I_m \otimes M) w_{\text{PE}} = g. \quad (\mathcal{C}')$$

Step 4. Solve the reduced system. Write $w_{\text{PE}} = (w_{\text{PE}}^{(1)}, \dots, w_{\text{PE}}^{(m)})$ and $g = (g_1, \dots, g_m)$ with blocks in \mathbb{R}^T . The Kronecker structure of $I_m \otimes M$ splits (\mathcal{C}') into m independent systems

$$M w_{\text{PE}}^{(i)} = g_i, \quad i = 1, \dots, m. \quad (\mathcal{C}_i)$$

Because $d_{\text{PE}} \geq T$, the rows of M are linearly dependent at worst; pick any right inverse M^+ (for instance the Moore–Penrose pseudoinverse). Then

$$w_{\text{PE}}^{(i)} := M^+ g_i \quad (i = 1, \dots, m)$$

solves each (\mathcal{C}_i) , and hence w_{PE} solves the full system (\mathcal{C}) .

Step 5. Conclusion. Compatibility holds for every choice of data $\{g_i\}$ once the frequencies obey $\omega_n = \frac{2\pi k_n}{L}$. \square

²least common multiple