Predicting the Unpredictable: Uncertainty-Aware Reasoning over Temporal Knowledge Graphs via Diffusion Process

Anonymous ACL submission

Abstract

Temporal Knowledge Graph (TKG) reasoning seeks to predict future incomplete facts leveraging historical data. While existing approaches have shown effectiveness in addressing the task through various perspectives, such as graph learning and logic rules, they are limited in capturing the indeterminacy in future events, particularly in the case of rare/unseen facts. To tackle the highlighted issues, we introduce a novel approach by conceptualizing TKG reasoning as a sequence denoising process for future facts, namely DiffuTKG. Concretely, we first encodes the historical events as the conditional sequence. Then we gradually introduce Gaussian noise to corrupt target facts during the forward process and then employ a transformer-based conditional denoiser to restore them in the reverse phase. Moreover, we introduce an uncertainty regularization loss to mitigate the risk of prediction biases by favoring frequent scenarios over rare/unseen facts. Empirical results on four real-world datasets show that DiffuTKG outperforms state-of-the-art methods across multiple evaluation metrics [1].

1 Introduction

Temporal Knowledge Graphs (TKGs) are dynamic, multi-relational structures that encapsulate the progression of events and knowledge in the real world, represented as quadruples (s, r, o, t), such as (Biden, meet, Zelensky, 2022-12-21). The reasoning tasks over TKGs are classified based on the temporal scope: interpolation involves inferring missing information within the observed timeframe, while extrapolation aims at predicting future events. Our research specifically focuses on extrapolation in TKGs, a domain that has more practical implications due to its forward-looking nature.

Existing studies (Trivedi et al., 2017; Jin et al., 2020; Li et al., 2021b) in TKG reasoning typically aggregate adjacent structure information and temporal information to derive the deterministic representations of entities and relations (Li et al., 2021a; Liu et al., 2023). These representations are subsequently applied within a scoring function, such as ConvTransE (Dettmers et al., 2018), to assess the likelihood of potential future facts (events).

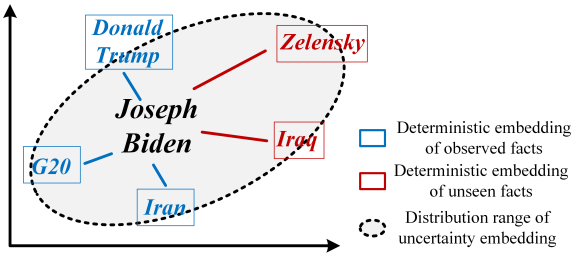

Figure 1: An example demonstrates how deterministic embeddings face challenges in managing uncertainty.

Despite the significant progress in TKG reasoning techniques, these deterministic methods exhibit inherent deficiencies when it comes to grappling with the uncertainties that arise from the unpredictable nature of future interactions (Jin et al., 2020; Sun et al., 2021) and the evolving understanding of temporal and structural relationships over time (Trivedi et al., 2017; Li et al., 2021b; Park et al., 2022a). This challenge is particularly evident in scenarios characterized by a scarcity of discriminative information, especially for facts with sparse or even no historical interactions (Chekol et al., 2017; Chen et al., 2019; Ji et al., 2021). These conventional approaches, which minimize the plausibility scores of unseen relation facts via the maximum likelihood objective, operate under the presumption that all unseen relation facts are false beliefs. As a result, they fail to capture the subtle uncertainty associated with these unseen facts.

To illustrate, consider the scenario depicted in Figure 1, *Biden* associated with only three historical facts and is anticipated to engage with the relatively rare or previously unseen facts *Iraq* and

[1] The source code is anonymized online at: https://anonymous.4open.science/r/DiffuTKG-049F

Zelensky in future scenarios. If we utilize deterministic embeddings derived from historical events to represent *Biden*, *Biden*'s position in the embedding space (mapped into a 2D map) may fall somewhere in the middle of *Donald Trump*, *G20*, and *Iran*. In such a setting, if predictions are based on the proximity within this embedding space, *Biden* is more likely to be forecasted to interact with *Donald Trump*, *G20*, and *Iran*, due to the closer embeddings. Furthermore, The widespread use of the maximum likelihood objective, such as cross-entropy loss, exacerbates prediction biases by favoring historically frequent scenarios over rare or unseen interactions (Zadeh and Schmid, 2021), thus hindering the model's adaptability to the unpredictable dynamics and emerging relationships inherent in real-world TKG scenarios.

To address these challenges, we propose DiffuTKG, a novel approach that reformulates TKG reasoning into a sequence prediction task by managing the inherent uncertainties through a sequence denoising method. In the training phase, DiffuTKG systematically transforms sequences of objects, relationships, and temporal intervals relevant to subject entities into a unified continuous representation. This process is then augmented by the strategic introduction of Gaussian noise, simulating the uncertain nature of future events (distribution ranges depicted in Figure 1). Subsequently, DiffuTKG harnesses a transformer-based framework for the denoising and accurate reconstruction of target entities, with the process intricately conditioned on both relational and temporal insights to mirror the intricate dynamics of TKGs structure.

Furthermore, DiffuTKG integrates an uncertainty regularization loss, which aids in distinguishing between seen and rare/unseen events, thereby enhancing the model's predictive clarity and reducing overfitting tendencies. During inference, DiffuTKG employs a reverse diffusion step initialized with sampled Gaussian noise to predict missing entities, subsequently refining these predictions based on calculated confidence scores. Empirical studies conducted on four benchmark datasets demonstrate the effectiveness of DiffuTKG. In summary, our main contributions are as follows:

- To the best of our knowledge, DiffuTKG is the first effort that introduces the diffusion process into TKG reasoning to explicitly manage dynamic and uncertain nature of future events via stochastic sequence denoising process.
- We introduce an uncertainty regularization loss to mitigate the risk of prediction biases, ensuring the model does not disproportionately favor frequently occurring historical scenarios over rare or unseen facts.
- Extensive experiments conducted on four real-world datasets demonstrate that DiffuTKG yields new state-of-the-art performance.

2 Our Approach

In this section, we introduce the details of our framework as shown in Figure 2. We first formulate the task definition of TKG reasoning as follow.

Definition 1 (**Temporal Knowledge Graph**) A temporal knowledge graph (TKG), denoted as $\mathcal{G}$, serves as a dynamic, multi-relational network of entities interconnected through time-stamped relations. This structure is conceptualized as a series of chronological KG snapshots, represented as $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{t-1}\}$. Each snapshot $\mathcal{G}_{t_i} \in \mathcal{G}$, encapsulates the facts at a specific time t_i, expressed as time-stamped quadruple (s, r, o, t_i), where $s, o \in \mathcal{E}$ are the subject and object entities, respectively, and $r \in \mathcal{R}$ signifies the relational fact connecting s and o. To facilitate a bi-directional comprehension of relationships within the TKG (Kazemi and Poole, 2018), the inverse quadruple (o, r^{-1}, s, t_i) is systematically appended to $\mathcal{G}$.

Definition 2 (**Temporal Knowledge Graph Reasoning**) The primary aim of TKG reasoning is to enable extrapolative entity prediction. Specifically, this entails predicting either the missing object entity in a future relation $(s, r, ?, t)$ or the absent subject entity $(?, r, o, t)$ utilizing historical TKG snapshots $\mathcal{G}_{t-L-1:t-1} = \{\mathcal{G}_{t-L}, \mathcal{G}_{t-L+1}, \ldots, \mathcal{G}_{t-1}\}$ spanning the preceding L timesteps.

2.1 TKG Reasoning as Sequence Prediction

Let $\mathcal{G}_{0:t-1}$ be historical TKG snapshots and $q_t = (s, r, o, t)$ be the query quadruple. To adopt the diffusion process in TKG reasoning, we reshape the task as that of sequence prediction, which involves predicting the missing entities in q_t by utilizing the historical events associated with the query subject s from $\mathcal{G}_{0:t-1}$. The historical event sequence related to s, sorted chronologically according to the timestamps is formally denoted as $Q_{0:n-1} = \{(s, r_0, o_0, t_0), \ldots, (s, r_i, o_i, t_i), \ldots, (s, r_{n-1}, o_{n-1}, t_{n-1})\}$[2], where $t_0 \leq t_i \leq$

[2]For brevity, we omit the superscript s in $Q^s_{0:n-1}$ for subject s.

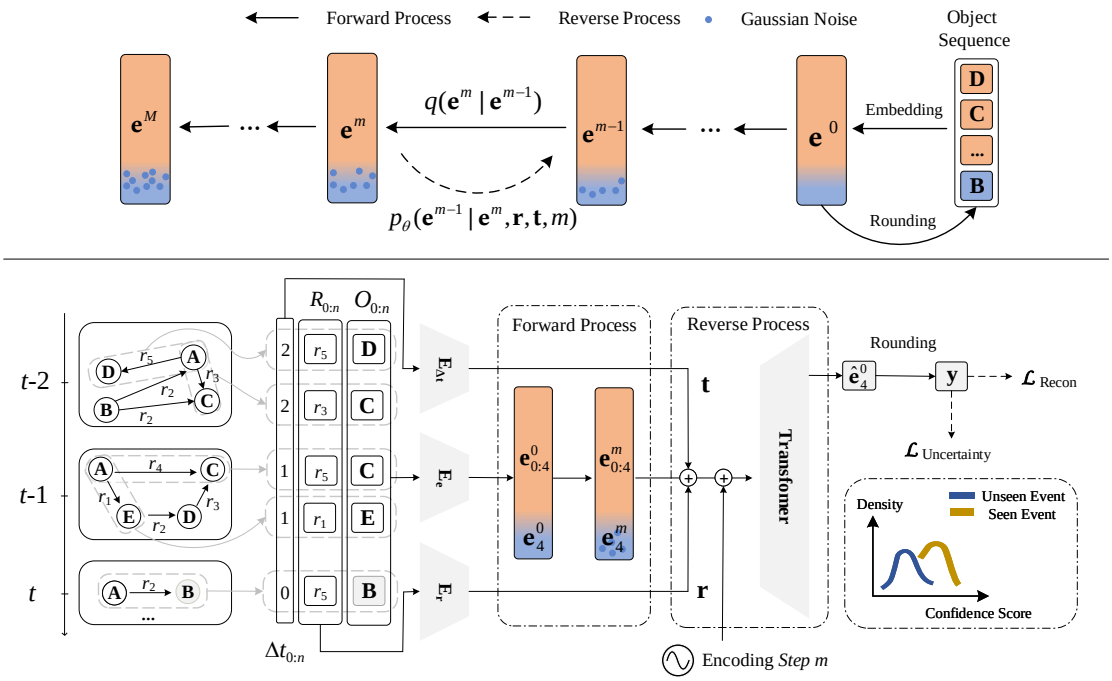

Figure 2: The upper part provides an overview of the diffusion process. We employ the color orange-red to symbolize historical objects associated with the query subject and cyan-blue to denote corresponding future objects. It's worth noting that noise is only added to the future object "B" in the forward process. The lower part illustrates the denoising training stage of DiffuTKG. In the figure, the TKGs at t-1 and t-2 represent the historical TKGs, while the TKG at t represents the future TKG. $\oplus$ denotes the element-wise addition operation.

$t_{n-1} < t$ and $n-1$ is the length of historical event sequence. Additionally, let $Q_{0:n-1} = \{S, R_{0:n-1}, O_{0:n-1}, T_{0:n-1}\}$. Here, $R_{0:n-1} = \{r_0, \ldots, r_{n-1}\}$ represents the sequence of relations in historical events, $O_{0:n-1} = \{o_0, \ldots, o_{n-1}\}$ denotes the sequence of objects in historical events, and $T_{0:n-1} = \{t_0, \ldots, t_{n-1}\}$ is the sequence of timestamps associated with historical events.

2.2 Denoising Training

The denoising training stage of DiffuTKG comprises three steps, focusing on reconstructing the missing object o while utilizing the historical event sequence $Q_{0:n-1}$ as conditioning factors.

Sequential Representation Learning In this phase, DiffuTKG is initially tasked with acquiring representations for objects and relations within $Q_{0:n} = \{Q_{0:n-1}, q_t\}$. Each object $o_i \in O_{0:n} = \{O_{0:n-1}, o\}$ is initially translated into its corresponding embedding vector $\mathbf{e}_i^0$ by the entity embedding matrix $\mathbf{E_e} \in \mathcal{R}^{d \times h}$. d is the number of entity types. Similarly, each relation $r_i \in R_{0:n} = \{R_{0:n}, r\}$ is projected into a continuous space using the relation embedding matrix $\mathbf{E_r}$. Additionally, we compute the time interval between every event and the queried event in $Q_{0:n}$, embedding them through $\mathbf{E_{\Delta t}} \in \mathcal{R}^{n \times h}$ for encoding temporal information. The projection process is denoted as:

$$\begin{aligned} \mathbf{e}^0 &= [\mathbf{e}_{0:n-1}^0; \mathbf{e}_n^0] \\ &= [\mathbf{E_e}(o_0); \mathbf{Emb_e}(o_1); \ldots, \mathbf{E_e}(o)], \\ \mathbf{r} &= [\mathbf{E_r}(r_0); \mathbf{E_r}(r_1); \ldots; \mathbf{E_r}(r)], \\ \mathbf{t} &= [\mathbf{E_{\Delta t}}(t); \mathbf{E_{\Delta t}}(t-1); \ldots; \mathbf{E_{\Delta t}}(0)], \end{aligned} \tag{1}$$

where $\mathbf{e}^0, \mathbf{r}, \mathbf{t} \in \mathbb{R}^{n \times h}$. $\mathbf{e}_{0:n-1}^0 \in \mathbb{R}^{(n-1) \times h}$ and $\mathbf{e}_n^0 \in \mathbb{R}^{1 \times h}$ represent the representations of objects in historical events and the representation of the target object, respectively, where h denotes the size of the hidden dimension. $[;]$ denotes the concatenation operation along the first dimension.

Forward Diffusion Process After obtaining the embedding of the object sequence $\mathbf{e}^0$, DiffuTKG specifically concentrates on introducing stochasticity incrementally to the target object $\mathbf{e}_n^0$. Consequently, the forward process is conceptualized as a Markov chain of Gaussian transitions:

$$q(\mathbf{e}_i^m \mid \mathbf{e}_i^0) = \begin{cases} \mathbf{e}_i^0 & \text{if } i < n \\ \sqrt{\bar{\alpha}_m}\mathbf{e}_i^0 + \sqrt{1-\bar{\alpha}_m}\epsilon & \text{if } i = n \end{cases} \tag{2}$$

The diffusion process extends over a specified range $m \in \{1, 2, \ldots, M\}$ and M marks the maximum number of forward steps. To regulate the added noises introduced by $1-\bar{\alpha}_m$, we use a linear noise schedule:

$$1-\bar{\alpha}_m = \delta \cdot \left[\alpha_{\min} + \tfrac{m-1}{M-1}\left(\alpha_{\max} - \alpha_{\min}\right)\right] \quad (3)$$

where the hyper-parameter $\delta \in [0, 1]$ controls the noise scales, and two hyper-parameters $\alpha_{\min} < \alpha_{\max} \in (0, 1)$ indicating the upper and lower bounds of the added noises.

Reverse Denoising Process In this phase, DiffuTKG undertakes the task of reconstructing the sequence of object entities from noise, with guidance from the temporal and relational characteristics of facts. More precisely, we introduce the encoded sequences of relations ($\mathbf{r}$) and time intervals ($\mathbf{t}$) to condition the denoising process as follows:

$$\begin{aligned} p_\theta\left(\hat{\mathbf{e}}^{m-1} \mid *\right) &= \mathcal{N}\left(\hat{\mathbf{e}}^{m-1}; \boldsymbol{\mu}_\theta\left(*\right), \boldsymbol{\Sigma}_\theta\left(*\right)\right), \\ \hat{\mathbf{e}}^{m-1} &= [\mathbf{e}^m_{0:n-1}; \hat{\mathbf{e}}^{m-1}_n], \end{aligned} \quad (4)$$

For brevity, we use the symbol "$*$" to represent $\{\hat{\mathbf{e}}^m, \mathbf{r}, \mathbf{t}, m\}$. $\hat{\mathbf{e}}^m$ is set to $\mathbf{e}^m$ at the first step of reverse process. Here, DiffuTKG adopts the architecture of a transformer encoder as f_θ to computing $\boldsymbol{\mu}_\theta\left(*\right)$ and $\boldsymbol{\Sigma}_\theta\left(*\right)$,which can be denoted as:

$$\begin{aligned} f_\theta(*) &= \text{Transformer}(\bar{\mathbf{e}}^m) = \hat{\mathbf{e}}^0, \\ \bar{\mathbf{e}}^m &= \hat{\mathbf{e}}^m + \mathbf{r} + \mathbf{t} + \text{Emb}_{\text{step}}(m). \end{aligned} \quad (5)$$

We incorporate step embeddings $\text{Emb}_{\text{step}}(\cdot)$ to manage the hidden representations at different noise levels (Gong et al., 2022).

2.3 Training Strategy

Reconstruction Loss Typically, diffusion models is trained using the Mean Square Error (MSE) loss (Shen et al., 2023), quantifying the difference between the original representation and the reconstructed one. However, MSE loss is recognized to be unstable in discrete space (Mahabadi et al., 2023; Dieleman et al., 2022). Hence, we opt for the dot product operation, which can stably quantify the distance between vectors:

$$\mathbf{y} = \text{Softmax}(f_\theta(\bar{\mathbf{e}}^m, m)_n \cdot (\mathbf{E}_{\mathbf{e}})^T) \quad (6)$$

where $f_\theta(\bar{\mathbf{e}}^m, m)_n \in \mathcal{R}^{1\times h}$ denotes the representation of the target object from $f_\theta(\bar{\mathbf{e}}^m, m)$ outputs. $(\cdot)^T$ is the matrix transposition operation and "$\cdot$" indicates the inner product operation. Consequently, to ensure conditional generation, we utilize a reconstruction loss function as follows:

$$\mathcal{L}_{\text{recon}} = -\sum_{i\in\{1,2,\ldots,d\}} g_i \log(\mathbf{y}_i), \quad (7)$$

where g_i represents the one-hot encoding of the i-th ground-truth object entity, and $\mathbf{y}_i$ is the predicted probability.

Uncertainty Loss Relying solely on traditional generation objectives can cause models to overfit to the labels of high-ranked outputs, a situation exacerbated in contexts of sparse and noisy data (Liu et al., 2020), potentially leading to incorrect assessments of both unseen and observed events. Additionally, due to the intrinsic generalization capabilities of diffusion models, their outputs might naturally contain clues for unseen events (Ma et al., 2024). To address this, we employ Multi-Layer Perceptrons (MLPs) to derive a confidence score from $\mathbf{y}$, providing supervisory signals for both unseen and seen events:

$$C(\mathbf{y}, F_{01}) = \text{MLP}(\text{Relu}(\text{MLP}(\mathbf{y} \otimes F_{01}))), \quad (8)$$

where $C(\mathbf{y}, F_{01}) \in \mathcal{R}^{1\times 2}$ denotes the confidence score, distinguishing the likelihood of an event being seen or unseen. The binary vector $F_{01} \in \mathcal{R}^{1\times d}$ records the historical occurrence of the event (s, r, o) before time t, with further details provided in Appendix D. We define P_{seen} as the set of logits for observed events and P_{non} for those unseen. To establish a distinct boundary between these, we introduce an uncertainty loss based on P_{seen} and P_{non}, encouraging DiffuTKG to assign higher confidence scores to features of seen events and lower scores to those of unseen events:

$$\begin{aligned} \mathcal{L}_{\text{uncertainty}} = &\mathbb{E}_{u\backsim P_{\text{seen}}}\left[-\log\frac{\exp^{-C(\mathbf{u},F_{01})/\tau}}{1+\exp^{-C(\mathbf{u},F_{01})/\tau}}\right] + \\ &\mathbb{E}_{v\backsim P_{\text{non}}}\left[-\log\frac{1}{1+\exp^{-C(\mathbf{v},F_{01})/\tau}}\right], \end{aligned} \quad (9)$$

where τ acts as a temperature coefficient, modulating the softness of the output probability distribution. Consequently, the overall training objective incorporates the reconstruction loss together with the uncertainty regularization loss, denoted as:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{uncertainty}}\,. \quad (10)$$

2.4 Sampling Inference

During inference, DiffuTKG samples Gaussian noise ϵ_n and applies the learned denoising model

f_θ for M reverse processes to denoise ϵ_n. The time complexity increases by M compared to training. To mitigate this computational overhead, we observe that f_θ is trained to directly predict $\hat{\mathbf{e}}^0$ based on any $\bar{\mathbf{e}}^m$ $(1 \leq m \leq M)$, so it can directly predict $\hat{\mathbf{e}}^0$ from $\bar{\mathbf{e}}^M$ without the need of the intermediate diffusion steps. Therefore, we design an efficient inference procedure by directly predicting $\hat{\mathbf{e}}^0$ from $\bar{\mathbf{e}}^M$:

$$\begin{aligned} \bar{\mathbf{e}}^M = \mathbf{e}^M + \mathbf{r} + \mathbf{t} = [\hat{\mathbf{e}}^M_{0:n}; \epsilon_n] + \mathbf{r} + \mathbf{t}, \\ \hat{\mathbf{e}}^0 = f_\theta(\bar{\mathbf{e}}^M, M). \end{aligned} \tag{11}$$

In line with the principles of ranking problems in graph reasoning (Jin et al., 2020), DIGM first computes the rank for each candidate entity using $\mathbf{y}$ from Equation (6). Then we calculate the confidence score c for the event features using Equation (8) and refine the ranking by dynamically incorporating prior frequency knowledge, similar to Liu et al. (2022a):

$$\mathbf{y} = \mathbf{y} + \lambda \times (\sigma(F) - F_{01}) \times \sigma(c), \tag{12}$$

where $F \in \mathcal{R}^{1 \times d}$ records the frequency of occurrences of the current event (s, r, o) before t, as detailed in Appendix D. σ denotes the softmax function. $\sigma(F)$ is employed to encourage an increase in the score of popular events, while "-$\sigma(F_{01})$" is used to suppress the occurrence of unseen events. The hyperparameter λ controls the effect of prior frequency knowledge.

3 Experiments

3.1 Datasets

We conduct the experimental evaluation on four TKG datasets to validate the effectiveness of our proposed model, which includes the ICEWS14, ICEWS05-15, ICEWS18 and GDELT datasets. The ICEWS series are from the Integrated Crisis Early Warning System (Boschee et al., 2015). GDELT is from the Global Database of Events, Language, and Tone (Leetaru and Schrodt, 2013). The data split strategy and data statistics are summarized in Appendix B.

3.2 Baseline Models

Fifteen typical exploration TKGR models are selected as the compared baselines, including RE-NET (Jin et al., 2020), RE-GCN (Li et al., 2021b), TANGO (Han et al., 2021b), xERTE (Han et al., 2021a), TiRGN (Li et al., 2022b), CEN (Li et al., 2022c), CENET (Xu et al., 2023), RETIA (Liu et al., 2023), HGLS (Zhang et al., 2023b), DaeMon (Dong et al., 2023), RPC (Liang et al., 2023a) , L2TKG (Zhang et al., 2023a), CluSTer (Li et al., 2021a), TITer (Sun et al., 2021), Tlogic (Liu et al., 2022b) and TECHS (Lin et al., 2023). We provide implementation details of baselines and DiffuTKG in Appendix C and D, respectively.

3.3 Evaluation Protocol

We assess our model's performance using standard metrics in the field: Mean Reciprocal Rank (MRR), Hits@1, and Hits@10. To ensure a fair comparison, we follow the experimental setup outlined by Li et al. (2021b), which includes the integration of ground truth historical data during multi-step inference. The results of our experiments are reported under a time-filtered setting, as detailed in (Dong et al., 2023; Zhang et al., 2023a).

3.4 Main Results

The comparative performance of various baseline models on the entity prediction task is detailed in Table 1, where the efficacy of denoising training in TKG reasoning is underscored by the results. Specifically, DiffuTKG exhibits significant improvements over the next best model, enhancing the Mean Reciprocal Rank (MRR) by absolute margins of 1.40%, 1.81%, and 2.67% in the ICEWS14, ICEWS18, and GDELT datasets, respectively. Notably, DiffuTKG demonstrates more pronounced performance gains on the GDELT dataset compared to the ICEWS datasets. This difference is attributed to the GDELT dataset's higher incidence of noisy data (Zhang et al., 2023a), which tends to obscure valuable discriminative information and leads to biased representations of entities. By incorporating uncertainty into entity representations, DiffuTKG effectively counters those scenarios, outperforming current state-of-the-art baselines. In the case of the ICEWS05-15 dataset, it includes a higher number of high-quality facts at each time, diminishing the necessity for uncertainty modeling. As a result, our model demonstrates limited improvement compared to state-of-the-art models in the ICEWS05-15 dataset.

3.5 Performance on Unseen Events

To further validate the capacity of DiffuTKG in capturing uncertainty information, we evaluate its performance on ICEWS datasets with unseen events

Table 1: Model performance (%) for the entity prediction task on ICEWS and GELDT datasets. The best results are highlighted in **bold** and the results of the second-best are underlined. The results marked with † are reproduced using their released code, those marked with * are from our reimplementation with default settings, and other results are retrieved from the original paper.

Models	ICEWS14			ICEWS05-15			ICEWS18			GEDLT		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
RE-NET	39.86	30.11	58.21	43.67	33.55	62.72	29.78	19.73	48.46	19.55	12.38	34.00
RE-GCN	42.00	31.63	61.65	48.03	37.33	68.51	32.62	22.39	52.68	19.69	12.46	33.81
TANGO	19.66	12.50	33.55	42.86	32.72	62.34	28.97	19.51	47.51	19.66	12.50	33.55
TITer	41.73	32.74	58.44	47.60	38.29	64.86	28.44	20.06	44.33	18.19	11.52	31.00
xERTE	40.79	32.70	57.30	46.62	37.84	63.92	29.31	21.03	46.48	19.45	11.92	34.18
TiRGN	43.81	33.49	63.50	49.84	39.07	70.11	33.58	23.10	54.20	21.67	13.63	37.60
CEN	42.20	32.08	61.31	-	-	-	31.50	21.70	50.59	-	-	-
Tlogic	43.04	33.56	61.23	46.97	36.21	67.43	29.82	20.54	48.53	-	-	-
TECHS	43.88	34.59	61.95	48.38	38.34	68.92	30.85	21.81	49.82	-	-	-
CENET	41.30	32.58	58.22	47.13	37.25	67.61	29.65	19.98	48.23	19.73	12.04	34.98
DaeMon	-	-	-	-	-	-	31.85	22.67	49.80	20.73	13.65	34.23
RPC	44.55	34.87	65.08	51.14	39.47	71.75	34.91	24.34	55.89	22.41	14.42	38.33
L2TKG *	45.89	34.63	68.47	52.42	40.09	75.86	31.63	21.17	53.01	20.16	12.49	35.83
HGLS †	47.11	35.87	70.61	47.17	36.83	68.89	30.18	20.63	50.23	19.87	12.19	35.43
RETIA †	46.20	35.39	68.70	52.29	40.33	74.18	34.86	24.10	56.96	-	-	-
DiffuTKG	**48.51**	**36.41**	**72.75**	**52.69**	**40.35**	**75.97**	**36.72**	**25.73**	**57.81**	**25.08**	**16.25**	**42.34**

Table 2: Performance of DiffuTKG, L2TKG, RETIA, CEN, and RE-GCN on predicting unseen events in terms of MRR and Hit@1 (%).

Models	ICEWS14		ICEWS18	
	MRR	Hit@1	MRR	Hit@1
RE-GCN	23.26	13.91	15.08	7.09
CEN	22.06	13.28	15.41	8.20
RETIA	24.17	14.67	16.62	9.08
L2TKG	23.88	14.35	16.48	8.84
DiffuTKG	**25.22**	**15.23**	**18.93**	**10.76**

that do not appear in the historical TKGs. The proportions of unseen events in the ICEWS datasets are presented in Table 4. We select four significant methods as comparative models, namely RE-GCN, CEN, RETIA, and L2TKG. The results presented in Table 2 indicate that DiffuTKG outperforms the baseline models. In comparison with other models, such as the SOTA model RETIA, our metrics have seen substantial relative improvements of 13.90% and 18.50% in ICEWS18. It's worth noting that the ICEW18 dataset contains a high proportion of unseen events (49.57%), indicating a high degree of sparsity in the occurrences of future events. It is evident that our network adeptly captures the uncertainty of event trends, especially in situations where uncertainty is pervasive within sparse datasets.

3.6 Ablation Studies

To verify the effectiveness of each module in DiffuTKG, ablation studies are carried out in Table 3. The first variant version "w/o $\mathbf{E_r}$" remove the relation embedding in f_θ. "w/o $\mathbf{E_{\Delta t}}$" means we remove the the embedding of time intervals in f_θ. "w/o $\mathcal{L}_{uncertainty}$" removes the uncertainty loss. And "$\mathcal{L}_{recon}$ as MSE" replaces the cross-entropy loss with the form of Mean Squared Error (MSE) for the reconstruction loss. We have the following observations: **(1)** the MRR values of "w/o $\mathbf{E_r}$" and "w/o $\mathbf{E_{\Delta t}}$" are much lower than that of DiffuTKG, which verifies the necessity of injecting temporal evolution and relation dependence into the denoising process; **(2)** "w/o $\mathcal{L}_{uncertainty}$" fails to leverage the complete generalized knowledge from the reconstruction representation, resulting in an overfitting issue. This leads to a relatively significant drop in reasoning performance, particularly on smaller datasets such as ICEWS14; **(3)** As anticipated, the model's reconstruction ability, trained through "$\mathcal{L}_{recon}$ as MSE", is unstable and adversely affects performance across the four datasets.

3.7 Sensitivity Analysis

We run our model with different important hyperparameters to explore weight impacts.

Figure 3 shows the changes in the performance of models with different lengths of the event sequence n, where small values would lead to great

Table 3: Ablation studies on all datasets in terms of MRR (%) with time-filter metrics.

Model	ICEWS14	ICEWS18	ICEWS05-15	GDELT
w/o $\mathbf{E_r}$	31.78	20.94	35.74	14.33
w/o $\mathbf{E_{\Delta t}}$	32.65	20.22	34.40	17.67
w/o $\mathcal{L}_{uncertainty}$	44.01	34.58	48.91	23.01
$\mathcal{L}_{recon}$ as MSE	39.87	27.89	46.78	15.91
DiffuTKG	**48.51**	**36.71**	**52.69**	**25.08**

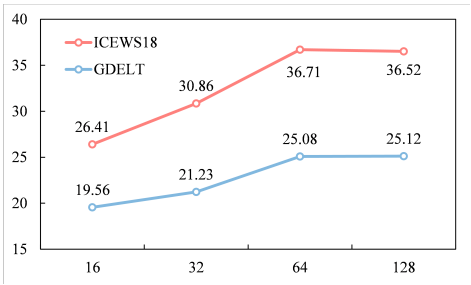

Figure 3: Performance of L2TKG under different length of the event sequence n in terms of MRR (%).

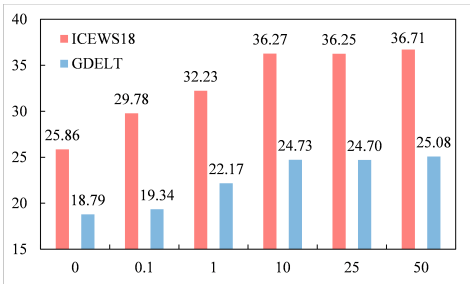

Figure 4: Performance of L2TKG under different noise scale δ in terms of MRR (%).

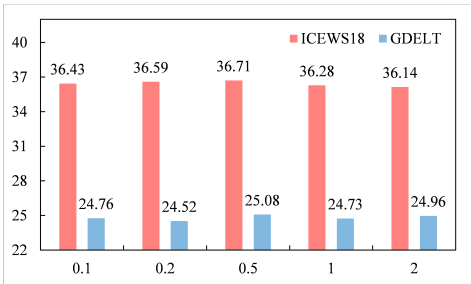

Figure 5: Performance of DiffuTKG under different temperature coefficient τ in terms of MRR (%).

performance decline. This is because fewer historical events lead to providing insufficient supervision signals for prediction, respectively. Nevertheless, an excessively long sequence of historical events can also result in information redundancy, thus limiting performance gains. So n is set to 64 for achieving optimal performance.

Different noise scales for the diffusion forward process are compared in Figure 4. As the noise scale increases, the performance first rises compared to training without noise ($s = 0$), verifying the effectiveness of denoising training. Furthermore, enlarging noise scales does not degrade performance, as the forward process only corrupts the target object and effectively retains event patterns in historical event sequences. Therefore, we can set $\delta \geq 20$, such as 50, to achieve satisfactory performance for all datasets.

Figure 5 demonstrates the impact of different temperature coefficients τ in $\mathcal{L}_{uncertainty}$. Setting the coefficient to a moderate value, generally 0.5, tends to yield the best result. It is worth noting that a smaller τ results in DiffuTKG placing more emphasis on events that are challenging to distinguish. Thus, carefully mining hard unseen events considerately prove functional for extrapolation reasoning on TKGs.

Figure 6 demonstrates that the model achieves optimal performance when λ is set to 2. Excessive or insufficient values for the hyperparameter can result in an imbalance of posterior and prior knowledge, leading to suboptimal results.

3.8 Inference Efficiency

To investigate the efficiency of our proposed model, we compare DiffuTKG with RETIA, xERTE, TiRCN and L2TKG in terms of inference time on the test set. Figure 7 illustrates that DiffuTKG is faster than other models. We attribute this to the fact that the model mainly consists of two linear attention layers, resulting in lower computational complexity. However, other models tend to be more time-consuming due to the inability to parallelize many computations, especially in RETIA. In summary, DiffuTKG ensures a significant improvement in time efficiency while delivering excellent extrapolation performance.

4 Related Work

4.1 TKG reasoning

TKG reasoning aims to predict facts in future events based on a sequence of observed historical facts. This task typically operates under two main scenarios: interpolation and extrapolation. In this work, our focus is primarily on the extrapolation aspect. Recently, The embedding-based approaches leverage temporal patterns (Jin et al., 2020; Li et al., 2021b) or structural information (Han et al., 2021b; Li et al., 2022d) to enhance prediction results. CEN (Li et al., 2022c) captures structure-variability evolutional patterns by a length-aware CNN. L2TKG

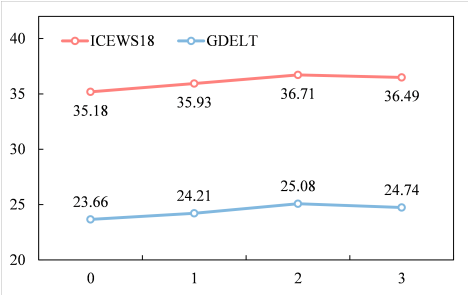

Figure 6: Performance of DiffuTKG under different λ values in terms of MRR (%).

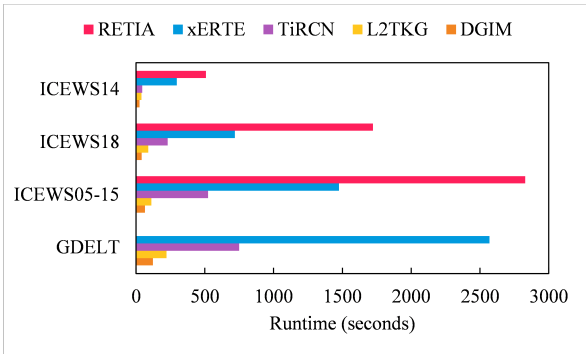

Figure 7: Runtime (seconds) comparison to some baselines. For ease of comparison, RETIA's inference times for ICEWS18 and ICEWS05-15 are scaled to one-fifth for comparison, while RETIA's data for GDELT is omitted as it is not applicable.

(Zhang et al., 2023a) exploits the intra-time relations between co-occurring entities and inter-time relations between entities that appear at different times. PRC (Liang et al., 2023b) further models the relational correlations in the intra-time information and periodic patterns in the inter-time interactions via two novel correspondence units. Considering the long-term dependencies among entities and relations, some works model the event time (Park et al., 2022b) and the long- and short-term entity and relation representations (Zhang et al., 2023b). DaeMon (Dong et al., 2023) and RETIA (Liu et al., 2023) focus on modeling the relation feature to adaptively capture the structure and temporal information. Some TKG reasoning methods leverage reward functions to enhance prediction results, such as the time-shaped reward (Sun et al., 2021) and beam-level reward (Li et al., 2021a). Ruled-base methods also are choices for TKG reasoning (Omran et al., 2019; Lin et al., 2023). Tlogic (Liu et al., 2022b) proposes a symbolic framework based on temporal logical rules extracted via temporal random walks. However, all of the aforementioned architectures overlook the uncertainty of future events, which is particularly common in events that occur rarely or never occur.

To tackle the above issues, DiffuTKG centers on a novel modeling paradigm from the perspective of sequence denoising generation. DiffuTKG is the first one to explore the utilization of the diffusion model on TKG reasoning, which infers future events from uncertainty in Gaussian noise.

4.2 Diffusion models on Discrete Data

Diffusion models (DMs) Sohl-Dickstein et al. (2015); Ho et al. (2020) have recently demonstrated the ability for high-quality generation across various domains, including image generation (Rombach et al., 2022; Ruiz et al., 2023) and audio generation (Borsos et al., 2023). Some efforts have sought to extend the applicability of continuous diffusion models into discrete spaces. Notably, Diffusion-LM (Li et al., 2022a) pioneers the adaptation of continuous diffusion models for text, incorporating an embedding step, a rounding step, and a dedicated training objective for embedding learning. Building upon this, DiffuSeq (Gong et al., 2022) introduces partial noise during the forward process, tailored for sequence-to-sequence tasks. Additionally, DiffusionNER (Shen et al., 2023) frames named entity recognition as a boundary-denoising diffusion process, effectively generating named entities from noisy spans. Despite the notable success of DMs in various domains, their application to TKG reasoning remains unexplored.

5 Conclusion

In this study, We introduce DiffuTKG, a novel paradigm that reconceptualizes TKG reasoning as a denoising diffusion process, tailored to address the inherent uncertainties within future facts. During the denoising training phase, we initiate the process by generating embeddings from historical data as conditional inputs. Following this, we methodically introduce Gaussian noise to the target entities, reflecting the uncertainty of future facts, and utilize a conditional denoising decoder for their accurate reconstruction. In addition to reconstruction loss, we incorporate an auxiliary loss aimed at reducing prediction biases, particularly those arising from an overemphasis on historically frequent scenarios at the expense of rare or previously unseen facts. Our empirical evaluations across various benchmark datasets confirm DiffuTKG's superior performance and efficiency in inference.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful discussion and feedback. This work is supported by (ANONYMIZED FOR DOUBLE BLIND REVIEW).

Limitations

In this section, we discuss the limitations of DiffuTKG. First, to maintain a simplified model, we opted not to increase complexity and directly input all historical events into the model without applying correlation-based filtering. However, this approach inevitably results in the inclusion of redundant information. Second, while the effectiveness of uncertainty loss has been demonstrated, the method of calculating the score has not been explored in depth, except for using simple nonlinear MLPs.

Ethics Statement

To ensure ethical considerations, we will provide a detailed description as follows:

1. All of the datasets used are collected and annotated in previous studies. The use of these datasets in our work does not involve any interaction or collection of individual privacy data.

2. Our work focuses on methodology studies and experiments. The results and models in our paper will not be used to harm or deceive any individuals or groups.

3. There are no potential conflicts of interest or ethical issues regarding financial support in the sponsors and funds of our research work.

References

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. ICEWS Coded Event Data.

Melisachew Wudage Chekol, Giuseppe Pirrò, Joerg Schoenfisch, and Heiner Stuckenschmidt. 2017. Marrying uncertainty and time in knowledge graphs. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, volume 31 of *AAAI'17*, page 88–94. AAAI Press.

Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. 2019. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3363–3370.

Tim Dettmers, Minervini Pasquale, Stenetorp Pontus, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, pages 1811–1818.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. 2022. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*.

Hao Dong, Zhiyuan Ning, Pengyang Wang, Ziyue Qiao, Pengfei Wang, Yuanchun Zhou, and Yanjie Fu. 2023. Adaptive path-memory network for temporal knowledge graph reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 2086–2094, Macao, China. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations*, pages 1–20, Kigali, Rwanda.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. 2021a. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. In *International Conference on Learning Representations*, pages 1–24.

Zhen Han, Zifeng Ding, Yunpu Ma, Yujia Gu, and Volker Tresp. 2021b. Learning neural ordinary equations for forecasting future links on temporal knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8352–8364, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pages 6840–6851, Online.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. 2020. Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6669–6683, Online. Association for Computational Linguistics.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kalev Leetaru and Philip A. Schrodt. 2013. Gdelt: Global data on events, location, and tone. *ISA Annual Convention*, 2:1–49.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022a. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 4328–4343, New Orleans, USA. Curran Associates, Inc.

Yujia Li, Shiliang Sun, and Jing Zhao. 2022b. Tirgn: Time-guided recurrent graph network with local-global historical patterns for temporal knowledge graph reasoning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2152–2158.

Zixuan Li, Saiping Guan, Xiaolong Jin, Weihua Peng, Yajuan Lyu, Yong Zhu, Long Bai, Wei Li, Jiafeng Guo, and Xueqi Cheng. 2022c. 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, Dublin, Ireland. Association for Computational Linguistics.

Zixuan Li, Zhongni Hou, Saiping Guan, Xiaolong Jin, Weihua Peng, Long Bai, Yajuan Lyu, Wei Li, Jiafeng Guo, and Xueqi Cheng. 2022d. HiSMatch: Historical structure matching based temporal knowledge graph reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7328–7338, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zixuan Li, Xiaolong Jin, Saiping Guan, Wei Li, Jiafeng Guo, Yuanzhuo Wang, and Xueqi Cheng. 2021a. Search from history and reason for future: Two-stage reasoning on temporal knowledge graphs. In *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 4732–4743, Online. Association for Computational Linguistics.

Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 2021b. Temporal knowledge graph reasoning based on evolutional representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 408–417, New York, NY, USA. Association for Computing Machinery.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. 2023a. 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*, SIGIR '23, page 1559–1568, New York, NY, USA. Association for Computing Machinery.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. 2023b. 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*, SIGIR '23, page 1559–1568, New York, NY, USA. Association for Computing Machinery.

Qika Lin, Jun Liu, Rui Mao, Fangzhi Xu, and Erik Cambria. 2023. TECHS: Temporal logical graph networks for explainable extrapolation reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1281–1293, Toronto, Canada. Association for Computational Linguistics.

Kangzheng Liu, Feng Zhao, Guandong Xu, Xianzhi Wang, and Hai Jin. 2022a. Temporal knowledge graph reasoning via time-distributed representation learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 279–288.

Kangzheng Liu, Feng Zhao, Guandong Xu, Xianzhi Wang, and Hai Jin. 2023. RETIA: relation-entity twin-interact aggregation for temporal knowledge graph extrapolation. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, pages 1761–1774. IEEE.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in neural information processing systems*, volume 33, pages 21464–21475, Vancouver, Canada.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. 2022b. 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, Vancouver, Canada.

Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Zhanhui Kang. 2024. Plug-in diffusion model for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada.

Rabeeh Karimi Mahabadi, Jaesung Tae, Hamish Ivison, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. 2023. Tess: Text-to-text self-conditioned simplex diffusion. *arXiv preprint arXiv:2305.08379*.

Pouya Ghiasnezhad Omran, Kewen Wang, and Zhe Wang. 2019. Learning temporal rules from knowledge graph streams. In *AAAI Spring Symposium Combining Machine Learning with Knowledge Engineering*.

Namyong Park, Fuchen Liu, Purvanshi Mehta, Dana Cristofor, Christos Faloutsos, and Yuxiao Dong. 2022a. Evokg: Jointly modeling event time and network structure for reasoning over temporal knowledge graphs. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 794–803, New York, NY, USA. Association for Computing Machinery.

Namyong Park, Fuchen Liu, Purvanshi Mehta, Dana Cristofor, Christos Faloutsos, and Yuxiao Dong. 2022b. Evokg: Jointly modeling event time and network structure for reasoning over temporal knowledge graphs. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 794–803, New York, NY, USA. Association for Computing Machinery.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, New Orleans, USA.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, Vancouver , Canada.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. DiffusionNER: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890, Toronto, Canada. Association for Computational Linguistics.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.

Haohai Sun, Jialun Zhong, Yunpu Ma, Zhen Han, and Kun He. 2021. TimeTraveler: Reinforcement learning for temporal knowledge graph forecasting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8306–8319, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. 2017. Know-evolve: deep temporal reasoning for dynamic knowledge graphs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3462–3471. JMLR.

Yi Xu, Junjie Ou, Hui Xu, and Luoyi Fu. 2023. Temporal knowledge graph reasoning with historical contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4765–4773.

Shekoufeh Gorgi Zadeh and Matthias Schmid. 2021. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3126–3137.

Mengqi Zhang, Yuwei Xia, Qiang Liu, Shu Wu, and Liang Wang. 2023a. Learning latent relations for temporal knowledge graph reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12617–12631, Toronto, Canada. Association for Computational Linguistics.

Mengqi Zhang, Yuwei Xia, Qiang Liu, Shu Wu, and Liang Wang. 2023b. Learning long- and short-term representations for temporal knowledge graph reasoning. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2412–2422, New York, NY, USA. Association for Computing Machinery.

A Diffusion Models for Discrete Data

The continuous diffusion model (DM) is a probabilistic model containing two Markov chains, mainly consisting of forward and reverse processes, which diffuse the data with pre-defined noise and reconstruct the desired sample from the noise (Ho et al., 2020). In this article, we center on DMs tailored for discrete data (Li et al., 2022a; Gong et al., 2022).

For the **forward diffusion** process of discrete data $\mathbf{w}$, an embedding step first transforms $\mathbf{w}$ into a continuous embedding $\mathbf{x}_0 \in \mathbb{R}^d$, parametrized by $q(\mathbf{x}_0|\mathbf{w}) = \mathcal{N}(\mathbf{x}_0, \mathrm{Emb}(\mathbf{w}), \beta_0\mathbf{I})$. In addition, $\mathrm{Emb}(\mathbf{w}) \in \mathbb{R}^d$ is an embedding function that maps each word to a vector in $\mathbb{R}^d$. Then the diffusion process corrupts $\mathbf{x}_0$ to obtain the latent variables $\mathbf{x}_{1:T}$ by gradually adding noise in T steps, where $\mathbf{x}_T$ is a standard Gaussian noise. The forward transition $\mathbf{x}_{t-1} \to \mathbf{x}_t$ can be attained by

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_{t-1}, \sqrt{1-\bar{\alpha}_t}\mathbf{I}) \\ &= \sqrt{\bar{\alpha}_t}\mathbf{x}_{t-1} + \sqrt{1-\bar{\alpha}_t}\epsilon, \end{aligned} \tag{13}$$

where $\mathcal{N}$ denotes the Gaussian distribution and $\epsilon \sim \mathcal{N}(0,1)$ is a random Gaussian noise. $\bar{\alpha}_t = \prod_{t'=1}^{t} \alpha_{t'} \in (0,1)$ controls the noise level at step $t \in \{0, 1, \ldots, T\}$.

The **reverse denoising** process takes the initial state $\mathbf{x}_T$ to reconstruct the original data $\mathbf{x}_0$ by learning from a neural network f_θ. The process can be formulated as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t); \Sigma_\theta(\mathbf{x}_t, t)) \tag{14}$$

where $\mu_\theta(\mathbf{x}_t, t)$ and $\Sigma_\theta(\mathbf{x}_t, t)$ represent the predicted parameterization of the mean and standard deviation, respectively, for $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}t)$, calculated by the function $f_\theta(\mathbf{x}_t, t)$. Finally, the rounding method, parametrized by $p_\theta(\mathbf{w}|\mathbf{x}_0) = \mathrm{Softmax}(\mathbf{x}_0)$, is employed to approximate or round the values to discrete representations. The corresponding training objective is

$$\begin{aligned} \mathcal{L}_{\text{simple}}^{\text{e2e}}(\mathbf{w}) = \mathop{\mathbb{E}}_{q_\phi(\mathbf{x}_{0:T}|\mathbf{w})} \left[\sum_{t=2}^{T} \left[\|\mathbf{x}_0 - f_\theta(\mathbf{x}_t, t)\|^2 \right] \right] + \\ \mathop{\mathbb{E}}_{q_\phi(\mathbf{x}_{0:1}|\mathbf{w})} \left[\|\mathrm{Emb}(\mathbf{w}) - f_\theta(\mathbf{x}_1, 1)\|^2 - \log p_\theta(\mathbf{w} \mid \mathbf{x}_0) \right]. \end{aligned} \tag{15}$$

The first expectation is to train the predicted model $f_\theta(\mathbf{x}_t, t)$ to approximate $\mathbf{x}_0$ from time step 2 to T. Empirically, it can effectively reduce rounding errors. The second expectation consists of two components: the first component aims to bring the predicted $\mathbf{x}_0$, closer to the embedding $\mathrm{Emb}(\mathbf{w})$, while the second component focuses on accurately rounding $\mathbf{x}_0$ to the text $\mathbf{w}$.

B Datasets

Followed by Li et al. (2021b)'s work, the data is split into training, validation, and test sets by 8:1:1 over the timeline. The detailed statistics of the datasets are presented in Table 4.

C Baselines

The comparison of TKG reasoning models with our work is presented as follows:

RE-NET (Jin et al., 2020) adopts RNN and RGCNs to capture the temporal and structural dependencies from entity sequences.

RE-GCN (Li et al., 2021b) proposes a novel Recurrent Evolution network based on Graph Convolution Network (GCN) to learns the evolutional representations of entities and relations at each timestamp by modeling the KG sequence recurrently

TANGO (Han et al., 2021b) proposes a multi-relational GCN to capture structural dependencies on TKGs and learns continuous dynamic representations using graph neural ordinary differential equations.

xERTE (Han et al., 2021a) reasons over query-relevant subgraphs of temporal KGs and jointly models the structural dependencies and the temporal dynamics.

TiRGN (Li et al., 2022b) employs a local recurrent graph encoder network to model the historical dependency of events at adjacent timestamps and utilizes a global history encoder network to gather repeated historical facts.

CEN (Li et al., 2022c) adopts a length-aware CNN to learn evolutional patterns of different lengths and explore online training strategy to deal with the problem of time-variability.

CENET (Xu et al., 2023) adopts contrastive learning to better guide the fusion of local and global historical information and enhance the ability to resist interference.

RETIA (Liu et al., 2023) evolutionally aggregates adjacent entity and relation features to produce relation embeddings on a twin hyperrelation subgraph sequence, thus spanning the message-passing gap.

Table 4: Dataset Statistics. $|\mathcal{V}|$ and $|\mathcal{R}|$ are the number of entity types and relation types. $|\mathcal{F}_{train}|$, $|\mathcal{F}_{valid}|$ and $|\mathcal{F}_{test}|$ are the numbers of fact triplets in training, validation, and test sets. The "Unseen Events" represents the proportions of queries encountering the dilemma of unseen events in the test set (%).

Datasets	$\|\mathcal{V}\|$	$\|\mathcal{R}\|$	$\|\mathcal{F}_{train}\|$	$\|\mathcal{F}_{valid}\|$	$\|\mathcal{F}_{test}\|$	Unseen Events
ICEWS14	6,869	230	377,430	36,588	28,644	47.63
ICEWS18	23,033	256	973,752	69,827	76.239	49.57
ICEWS05-15	10,094	251	1,430,389	171,518	156,695	31.61
GDELT	7,691	240	1,734,399	238,765	305,241	35.07

HGLS (Zhang et al., 2023b) transforms the TKG sequence into a global graph to explicitly associate historical entities at different times.

DaeMon (Dong et al., 2023) adaptively captures the temporal path information between query subject and object candidates across time by utilizing historical structural and temporal characteristics while considering the query feature.

RPC (Liang et al., 2023a) sufficiently mines the information underlying the Relational correlations and Periodic patterns via two novel Correspondence units.

L2TKG (Zhang et al., 2023a) exploits the intra-time and inter-time latent relations to alleviate the problem of missing associations in TKG reasoning.

CluSTer (Li et al., 2021a) learns a beam search policy via reinforcement learning (RL) to induce multiple clues from historical facts and adopts a GCN-based sequence method to deduce answers from clues.

TITer (Sun et al., 2021) navigates through TKG historical snapshots and searches for the temporal evidence chain to locate the target object.

Tlogic (Liu et al., 2022b) generates answers by applying rules to observed events before the query timestamp and scores the answer candidates relying on the rules' confidences and time differences.

TECHS (Lin et al., 2023) integrates propositional and first-order reasoning in a logical decoder to achieve explainability.

D Implementation Details

Hyperparameter settings We utilize the Adam optimizer with a learning rate set to 0.001 and l_2 regularization set to 1e-5. The number of training epochs is set to 100. Besides, the noise scale δ, the noise lower bound α_{min}, the noise upper bound α_{max} are 50,1e-2, respectively, with a total diffusion step T of 200. The length of historical TKGs denoted as L, is set to 64 for all datasets. The hidden size for entities and relations, denoted as h, is fixed at 200 for all datasets. The layer numbers of the transformer encoder are 2 for all datasets. The dropout rate is 0.2 for all datasets. The temperature coefficient τ is set to 0.5 across all datasets, and the scale parameter are searched in 2,3,4 for all datasets.

We report a statistically significant improvement ($p < 0.05$) based on the bootstrap paired t-test in our experimental results. The computational experiments in Section 3.8 are conducted on NVIDIA Tesla V100 (32G). Other experiments are conducted on NVIDIA Tesla A100 (80G).

Calculation Method for Frequency Information For the query (s, r, o, t), we store the event frequency using a sparse matrix $MF \in \mathcal{R}^{d \cdot w \times d}$, where w is the number of relations. Each row is represented as the vector $F = MF^{(s,r)} \in \mathcal{R}^d$, counting the number of occurrences. The multi-heat vector F_{01} is derived by converting F, where occurrences are recorded as 1, and the rest are set to 0.