

SHAPE: Structural Hypergraph Augmentation with Personalized Event Memory for Temporal Knowledge Graph Reasoning

Qing Li*, Jialin Lv, Jingjing Song, Duan fei, Zhongwen Guo

School of Computer Science, Northwestern Polytechnical University, Xi'an, China

{qingli@, nwpu_lvjialin@mail., songjingjing@mail.,
2025263042@mail., zhongwenguo@mail.}nwpu.edu.cn

Abstract

Temporal knowledge graph (TKG) reasoning aims to predict future events from historical facts. A key challenge for graph-based TKG models is to simultaneously capture higher-order structural dependencies and identify query-relevant temporal evidence, yet existing methods are largely limited to pairwise interactions or query-agnostic history aggregation. To address this, we propose SHAPE (Structural Hypergraph Augmentation with Personalized Event Memory), a plug-and-play enhancement framework for graph-based TKG models that strengthens the backbone at both structural and temporal levels without modifying its architecture. Specifically, Dual Hypergraph Pre-Enhancement(DHPE) constructs learnable entity and relation prototype hypergraphs with cross-view information exchange to inject higher-order structural priors before temporal encoding, while Query-Denoised Event Memory(QDEM) retrieves query-relevant recent events and converts them into subject enhancement and candidate-level scoring bias via structure-aware expert gating. Extensive experiments on three benchmarks with two heterogeneous graph backbones verify the effectiveness of SHAPE.

1 Introduction

Temporal knowledge graphs (TKGs) organize dynamic facts as timestamped quadruples (s, r, o, t) , where entities s, o are connected by relation r at discrete time t . Reasoning over TKGs aims to predict future events from historically observed facts and supports applications such as geopolitical forecasting (Bosch et al., 2015), financial risk assessment (Deng et al., 2020), and recommendation (Wang et al., 2019). The central challenge is to model both the structural dependencies inside each snapshot and the temporal dependencies across snapshots.

*Corresponding author.

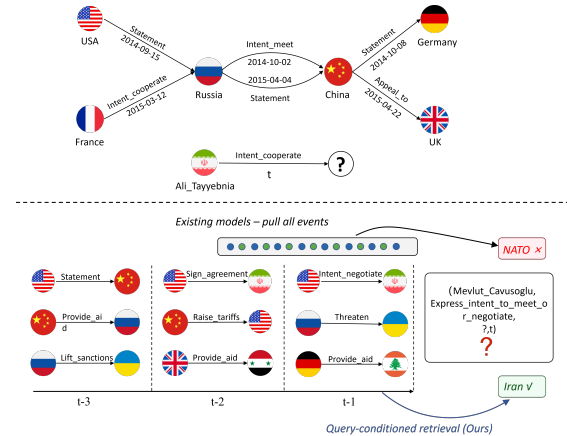


Figure 1: Illustration of the two problems of TKG reasoning task.

Recent methods have substantially improved TKG extrapolation by combining graph encoders with temporal models. RE-GCN (Li et al., 2021) couples relation-aware graph convolution with gated recurrence. TiRGN (Li et al., 2022) augments decoding with global historical vocabularies. HGLS (Zhang et al., 2023b), L2TKG (Zhang et al., 2023a), and LogCL (Chen et al., 2024) further strengthen long-range or query-aware history modeling. Despite this progress, two important limitations remain.

Limitation 1: Weak higher-order structural priors. Most existing encoders propagate messages only along pairwise edges, which captures local connectivity but cannot explicitly model higher-order semantic groups shared by multiple entities and relations. In ICEWS-style data, for example, countries such as the *United States* and *France* may exhibit similar diplomatic behaviors, yet such shared roles are difficult to recover from fragmented pairwise paths alone. As illustrated in the upper part of Figure 1, their local interactions with Russia are observable, but their common semantic role is not explicitly encoded.

Limitation 2: Query-agnostic use of recent history. Many decoder-side history modules aggre-

gate recent events with limited query conditioning, so irrelevant contemporaneous events may be mixed into the representation. For instance, a diplomatic query can be distracted by nearby economic or humanitarian events, weakening the supervision signal for the current reasoning target. The lower part of Figure 1 shows such a case, where query-relevant retrieval is polluted by unrelated timeline events.

To address both limitations within a single framework, we propose SHAPE (Structural Hypergraph Augmentation with Personalized Event Memory). SHAPE enhances a backbone at both ends of the reasoning pipeline: before temporal encoding, *Dual Hypergraph Pre-Enhancement* (DHPE; §3.3) learns higher-order structure through entity and relation prototype hypergraphs with a cross-view semantic bridge; after temporal encoding, *Query-Denoised Event Memory* (QDEM; §3.4) retrieves query-relevant recent events, refines the subject representation, and produces a candidate-level bias via structure-aware expert gating (SAEG). Since SHAPE only modifies encoder inputs and decoder-side representations, it can be plugged into existing TKG backbones without changing their internal architectures. Our main contributions are as follows:

- We propose DHPE, a dual hypergraph pre-enhancement module that learns entity and relation prototype hypergraphs end-to-end and injects higher-order structural priors before temporal encoding.
- We introduce QDEM, a query-conditioned memory module that performs denoised recent-event retrieval and produces both subject enhancement and candidate-level bias through the proposed SAEG mechanism.
- We show that SHAPE consistently improves two heterogeneous backbones, HGLS and LogCL, and achieves the best MRR on all three evaluated ICEWS benchmarks. Extensive ablation, sensitivity, efficiency, and sparsity analyses further validate the effectiveness of the design.

2 Related Work

TKG Reasoning. Early work extends static KG embeddings with time-aware representations (Dasgupta et al., 2018; Lacroix et al., 2020; Jain et al., 2020; Xu et al., 2021). For extrapolation, RE-NET (Jin et al., 2020) and RE-GCN (Li et al., 2021) combine graph encoders with recurrent tem-

poral updates; CyGNet (Zhu et al., 2021) and CENET (Xu et al., 2023) introduce copy and contrastive retrieval mechanisms to exploit long-term global history; TiRGN (Li et al., 2022) further models repetitive and cyclical historical patterns. xERTE (Han et al., 2021), TITer (Sun et al., 2021), and TLogic (Liu et al., 2022) retrieve query-relevant subgraphs or temporal rules for interpretable reasoning. HGLS (Zhang et al., 2023b) and LogCL (Chen et al., 2024) capture long- and short-term history via global graph construction and contrastive learning, while CognTKE (Chen et al., 2025) and DiMNet (Dong et al., 2025) further strengthen query-aware history selection and adaptive temporal evolution. Despite their diversity, most methods rely on pairwise propagation without a higher-order structural prior, and their decoder-side history mechanisms offer limited query-conditioned filtering.

Hypergraph and Memory-Augmented Reasoning. Hypergraph neural networks (Feng et al., 2019; Chien et al., 2022; Yadati et al., 2024; Huang et al., 2026) generalize GNNs to group-wise interactions. DHE-TKG (Liu et al., 2024) and MGESL (Shi et al., 2024) apply dynamic and multi-granularity hypergraph convolution to TKG forecasting, yet both build hyperedges via heuristic co-occurrence rules and restrict structural priors to the encoder side. On the memory side, CyGNet (Zhu et al., 2021), TiRGN (Li et al., 2022), HGLS (Zhang et al., 2023b), and LogCL (Chen et al., 2024) aggregate historical signals at decoding, while MemoTime (Tan et al., 2026) couples explicit retrieval with a self-evolving experience memory for LLM-grounded question answering. None of these methods coordinates query-conditioned memory scoring with a learnable higher-order structural prior.

3 Method

3.1 Preliminaries

A temporal knowledge graph $\mathcal{G} = \{(s, r, o, t)\}$ consists of timestamped facts, where $s, o \in \mathcal{E}$ are entities, $r \in \mathcal{R}$ is a relation, and $t \in \mathcal{T}$ is a discrete timestamp. We denote the snapshot at time t by $G_t = \{(s, r, o) \mid (s, r, o, t) \in \mathcal{G}\}$. Given a query $(s, r, ?, t)$ and historical snapshots $\{G_1, \dots, G_{t-1}\}$, the TKG extrapolation task aims to rank all candidate entities $o' \in \mathcal{E}$ such that the ground-truth object is ranked as high as possible. Following standard practice (Li et al., 2021,

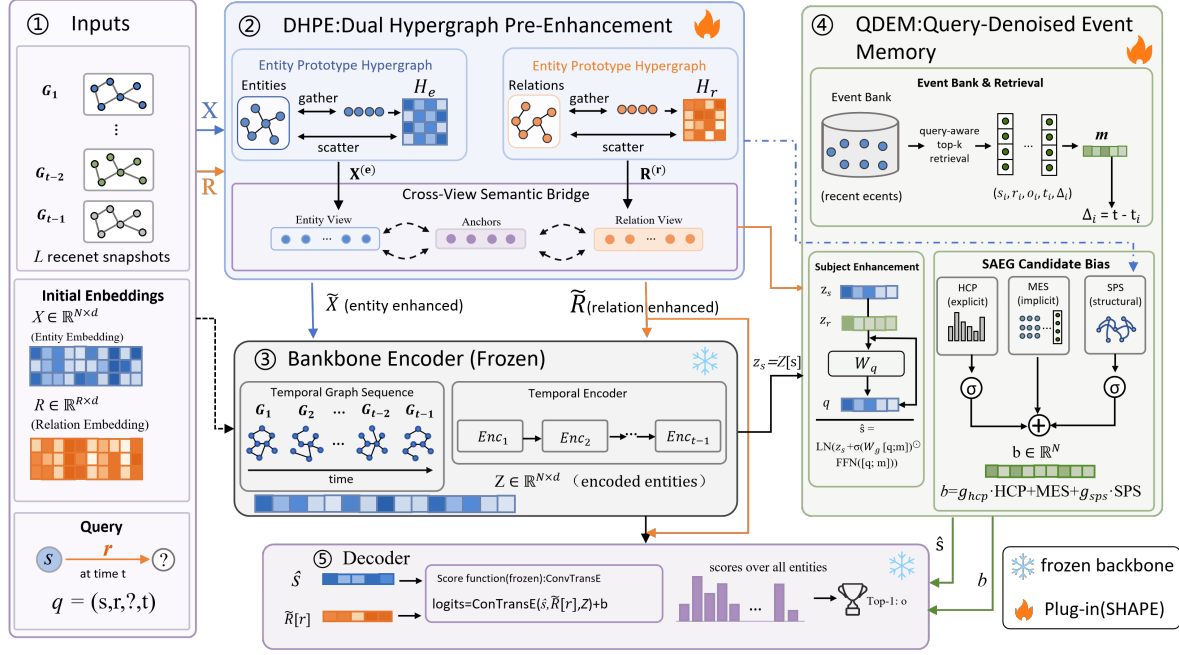


Figure 2: Overall architecture of SHAPE. DHPE is inserted *before* the backbone’s temporal encoder to produce structure-enhanced embeddings $\tilde{\mathbf{X}}, \tilde{\mathbf{R}}$. QDEM is inserted *after* the encoder to produce a query-specific enhanced subject $\hat{s}$ and a candidate scoring bias $\mathbf{b}$. The backbone encoder remains entirely unmodified.

2022; Chen et al., 2024), we add inverse facts (o, r^{-1}, s, t) to treat object and subject prediction symmetrically. We write $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{R} \in \mathbb{R}^{R \times d}$ for the entity and relation embedding matrices.

A graph-based TKG backbone typically follows an *embed* $\rightarrow$ *encode* $\rightarrow$ *decode* pipeline: the embed step initializes $\mathbf{X}, \mathbf{R}$; the encode step processes historical snapshots to produce updated entity representations $\mathbf{Z} \in \mathbb{R}^{N \times d}$; and the decode step scores each candidate against the query. SHAPE exploits this shared interface and is therefore applicable to diverse backbones without modifying their architecture.

3.2 Overview

SHAPE is designed as an enhancement framework. Its key design choice is to insert one module at each end of the standard TKG pipeline introduced in §3.1, while keeping the backbone encoder unchanged and reusing existing temporal modeling. The two modules address the two limitations identified in §1: weak higher-order structural priors at the embedding level and query-agnostic use of recent history at the decoding level. As shown in Figure 2, DHPE prepares structure-aware inputs for the encoder, whereas QDEM personalizes the decoder output for the current query.

3.2.1 Pre-encoder structural enhancement.

The first module, Dual Hypergraph Pre-Enhancement (DHPE; §3.3), replaces the raw entity and relation lookup tables $\mathbf{X}, \mathbf{R}$ with structure-aware embeddings $\tilde{\mathbf{X}}, \tilde{\mathbf{R}}$. It learns entity and relation prototype hypergraphs end to end and exchanges information between the two views through a cross-view semantic bridge. DHPE also outputs the entity-side soft prototype-membership matrix, which is passed to QDEM as a structural signal for candidate scoring.

3.2.2 Post-encoder query personalization.

The second module, Query-Denoised Event Memory (QDEM; §3.4), operates on the encoder output $\mathbf{Z}$ and the relation matrix $\tilde{\mathbf{R}}$. Rather than using the full recent-event window indiscriminately, it maintains a memory bank of recent facts and performs query-conditioned top- k retrieval. It then produces an enhanced subject representation $\hat{s}$ for candidate scoring and an additive bias $\mathbf{b} \in \mathbb{R}^N$ that combines historical prior, memory–entity similarity, and the prototype-membership signal from DHPE.

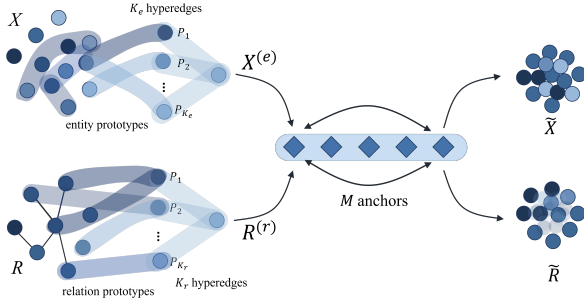


Figure 3: DHPE constructs entity and relation prototype hypergraphs via learnable prototypes, performs gather-scatter convolution, and links the two views through a cross-view semantic bridge.

3.2.3 Pipeline.

Putting these two modules together, the SHAPE pipeline can be summarized as

$$\begin{cases} (\tilde{\mathbf{X}}, \tilde{\mathbf{R}}) = \text{DHPE}(\mathbf{X}, \mathbf{R}), \\ \mathbf{Z} = \text{Encoder}(\tilde{\mathbf{X}}, \tilde{\mathbf{R}}, G_1, \dots, G_{t-1}), \\ (\hat{\mathbf{s}}, \mathbf{b}) = \text{QDEM}(\mathbf{Z}, \tilde{\mathbf{R}}, \mathcal{H}, q), \\ \text{logits} = \text{Decoder}(\hat{\mathbf{s}}, \mathbf{Z}) + \mathbf{b}, \end{cases} \quad (1)$$

where $\mathcal{H}$ denotes the recent event bank used by QDEM and $q = (s, r)$ is the current query. The backbone encoder appears in Eq. (1) as a black box: SHAPE only transforms its inputs and post-processes its outputs, and does not require access to the encoder’s internal modules or temporal modeling assumptions.

3.3 Dual Hypergraph Pre-Enhancement (DHPE)

DHPE addresses Limitation 1 by discovering higher-order semantic groups through learnable prototype hypergraphs and linking entity and relation views via a cross-view bridge, as illustrated in Figure 3. It contains three sub-modules.

3.3.1 Entity Prototype Hypergraph.

We define K_e learnable entity prototype vectors $\mathbf{P}_e \in \mathbb{R}^{K_e \times d}$, each representing a latent semantic group (e.g., sovereign states, political leaders). Before computing entity-prototype membership, $\mathbf{P}_e$ is first modulated by global relation context through a scaled dot-product affinity and an additive residual update, making the prototypes relation-aware from the outset:

$$\begin{aligned} \mathbf{P}'_e &= \mathbf{P}_e + \mathbf{A}_{\text{rp}}^\top \mathbf{R} \mathbf{W}_r, \\ \mathbf{A}_{\text{rp}} &= \text{softmax}\left(\frac{\mathbf{R} \mathbf{P}_e^\top}{\sqrt{d}}\right). \end{aligned} \quad (2)$$

where the softmax in $\mathbf{A}_{\text{rp}} \in \mathbb{R}^{R \times K_e}$ is taken column-wise so that each prototype receives a normalized distribution over relations. The modulated prototypes $\mathbf{P}'_e$ are then used to compute a soft incidence matrix via multi-head cross-attention with a learnable temperature τ :

$$\mathbf{H}_e = \text{softmax}\left(\frac{\text{mean}_{\text{head}}[\mathbf{X} \mathbf{W}_q (\mathbf{P}'_e \mathbf{W}_k)^\top]}{\tau \sqrt{d_h}}\right), \quad (3)$$

where $d_h = d/h$ is the per-head dimension and τ is clamped to $[0.1, +\infty)$. The resulting $\mathbf{H}_e \in \mathbb{R}^{N \times K_e}$ is a fully differentiable soft incidence matrix, where each row encodes an entity’s membership distribution over prototypes. An optional top- k sparsification mask retains only the k strongest memberships per entity to improve efficiency.

Given $\mathbf{H}_e$, we perform hypergraph convolution via the gather-scatter paradigm (Feng et al., 2019). The gather step aggregates entity features into prototype-level representations and refines them with a residual feed-forward block:

$$\tilde{\mathbf{Z}}_e = \mathbf{D}_e^{-1} \mathbf{H}_e^\top \mathbf{X}, \quad \mathbf{Z}_e = \tilde{\mathbf{Z}}_e + \text{FFN}(\tilde{\mathbf{Z}}_e), \quad (4)$$

where $\mathbf{D}_e = \text{diag}(\mathbf{H}_e^\top \mathbf{1})$ is the prototype-degree matrix. The scatter step then distributes the prototype features back to entities:

$$\mathbf{X}_{\text{agg}}^{(e)} = \text{ReLU}\left(\text{LN}\left(\mathbf{D}_v^{-\frac{1}{2}} \mathbf{H}_e \mathbf{Z}_e \mathbf{W}_v\right)\right), \quad (5)$$

where $\mathbf{D}_v = \text{diag}(\mathbf{H}_e \mathbf{1})$ is the vertex-degree matrix and $\mathbf{W}_v \in \mathbb{R}^{d \times d}$ projects prototype features. Finally, a gated residual balances the hypergraph-enhanced representation with a skip projection $\mathbf{X} \mathbf{W}_s$ of the original embedding ($\mathbf{W}_s \in \mathbb{R}^{d \times d}$ is a learnable skip projection):

$$\mathbf{g}_e = \sigma\left([\mathbf{X}_{\text{agg}}^{(e)}; \mathbf{X} \mathbf{W}_s] \mathbf{W}_g\right), \quad (6)$$

$$\mathbf{X}^{(e)} = \mathbf{g}_e \odot \mathbf{X}_{\text{agg}}^{(e)} + (1 - \mathbf{g}_e) \odot \mathbf{X} \mathbf{W}_s, \quad (7)$$

where $\mathbf{W}_g \in \mathbb{R}^{2d \times d}$ produces the per-dimension gate.

3.3.2 Relation Prototype Hypergraph.

A symmetric module operates on relation embeddings with K_r learnable relation prototypes $\mathbf{P}_r \in \mathbb{R}^{K_r \times d}$. The incidence matrix $\mathbf{H}_r \in \mathbb{R}^{R \times K_r}$ is computed analogously to Eq. (3) but uses $\mathbf{P}_r$ directly without prototype modulation, since relations are now the nodes rather than the modulators. To still inject cross-side information, the

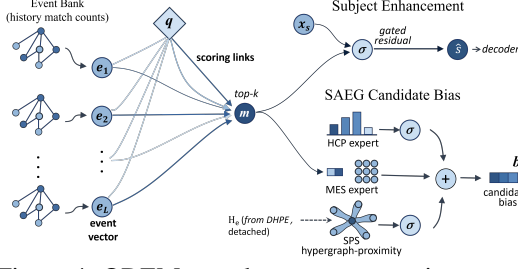


Figure 4: QDEM encodes recent events into a memory bank, performs query-conditioned top- k retrieval for subject enhancement, and generates a candidate scoring bias via three parallel signals (HCP, MES, SPS) fused by structure-aware expert gating.

entity-context vector $\bar{\mathbf{x}} = \text{mean}(\mathbf{X}^{(e)})$ is added to the gathered hyperedge features $\mathbf{Z}_r$ (Eq. (4), with $\mathbf{H}_r, \mathbf{R}$ in place of $\mathbf{H}_e, \mathbf{X}$), providing type-constraint priors. The same scatter step and gated residual then produce the enhanced relation embeddings $\mathbf{R}^{(r)}$.

3.3.3 Cross-View Semantic Bridge.

The entity and relation hypergraphs operate independently so far. To enable inter-view information flow, we introduce M learnable anchor vectors $\mathbf{B} \in \mathbb{R}^{M \times d}$ as a compact information bottleneck. Concretely, the anchors first absorb entity-view information via multi-head attention (MHA) and then broadcast it to the relation view; for instance, the entity-side cross-view context is computed as

$$\begin{aligned} \mathbf{X}^{\text{cv}} &= \text{MHA}(\mathbf{X}^{(e)}, \\ &\quad \text{MHA}(\mathbf{B}, \mathbf{R}^{(r)}, \mathbf{R}^{(r)}), \\ &\quad \text{MHA}(\mathbf{B}, \mathbf{R}^{(r)}, \mathbf{R}^{(r)})). \end{aligned} \quad (8)$$

and the relation-side context $\mathbf{R}^{\text{cv}}$ is obtained symmetrically by swapping the roles of $\mathbf{X}^{(e)}$ and $\mathbf{R}^{(r)}$. Because all attention paths go through the M anchors, the overall cost is $O(M(N + R))$, independent of $N \times R$. The two cross-view contexts are then fused with their own views through gated residuals (same form as Eq. (7)) and refined by a position-wise FFN with LayerNorm, yielding the final DHPE outputs $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{R}}$.

3.4 Query-Denoised Event Memory (QDEM)

QDEM addresses Limitation 2 by performing query-conditioned retrieval from recent history, as shown in Figure 4. It operates on the post-encoder entity matrix $\mathbf{Z}$ and the relation matrix $\tilde{\mathbf{R}}$ from the SHAPE pipeline (Eq. (1)); we write $\mathbf{z}_s$ for the row of $\mathbf{Z}$ corresponding to the query subject and $\mathbf{z}_r$ for the row of $\tilde{\mathbf{R}}$ corresponding to the query relation.

QDEM produces two outputs: an enhanced subject embedding $\hat{\mathbf{s}}$ and a candidate scoring bias $\mathbf{b}$.

3.4.1 Event Bank and Query-Conditioned Retrieval.

Given the L most recent snapshots, each historical fact (s_i, r_i, o_i, t_i) is encoded into a dense event vector $\mathbf{e}_i$ by concatenating the encoder-side subject, relation, object, and a learnable temporal position embedding indexed by the recency offset $\Delta_i = t - t_i$, projecting them through an FFN, and adding a role-direction term computed from $\mathbf{z}_{o_i} - \mathbf{z}_{s_i}$. The bank is augmented with the inverse fact $(o_i, r_i^{-1}, s_i, t_i)$ so that each original event contributes from both directions. For a query $q = (s, r, ?, t)$, we compute a biased relevance score for each event:

$$\begin{aligned} a_i &= \frac{\mathbf{q}^\top \mathbf{W}_k \mathbf{e}_i}{\sqrt{d}} - \lambda \frac{\Delta_i}{L} \\ &\quad + b_s \mathbb{I}[s_i = s] + b_r \mathbb{I}[r_i = r], \quad i \in \mathcal{H}. \end{aligned} \quad (9)$$

Here $\mathcal{H}$ is the index set of all events in the bank, $\mathbf{q} = \mathbf{W}_q[\mathbf{z}_s; \mathbf{z}_r]$ is the query vector, $\lambda \geq 0$ is a learnable recency decay (parameterized via soft-plus to ensure non-negativity), and b_s, b_r are learnable bias terms. Events whose forward or inverse form matches neither the query subject nor relation are masked out before top- k selection. Let $\mathcal{T}_k(q)$ denote the retained event indices for query q . We normalize the retained scores only within this top- k set:

$$\alpha_j = \frac{\exp(a_j)}{\sum_{\ell \in \mathcal{T}_k(q)} \exp(a_\ell)}, \quad j \in \mathcal{T}_k(q). \quad (10)$$

The memory context is then obtained by a weighted sum of the retained event values:

$$\mathbf{m} = \sum_{j \in \mathcal{T}_k(q)} \alpha_j \mathbf{W}_v \mathbf{e}_j. \quad (11)$$

3.4.2 Subject Enhancement.

The memory context $\mathbf{m}$ is fused with the encoder-side subject embedding through a gated residual to produce the enhanced subject:

$$\hat{\mathbf{s}} = \text{LN}(\mathbf{z}_s + \sigma(\mathbf{W}_g[\mathbf{q}; \mathbf{m}]) \odot \text{FFN}([\mathbf{q}; \mathbf{m}])). \quad (12)$$

$\hat{\mathbf{s}}$ replaces the encoder-side subject embedding for decoding, injecting query-specific historical context into the prediction.

3.4.3 Candidate Scoring with Structure-Aware Expert Gating (SAEG).

The candidate bias $\mathbf{b} \in \mathbb{R}^N$ fuses three complementary scoring signals, each acting on the same candidate set but exploiting a different source of evidence. **HCP** (Historical Candidate Prior) is an explicit, recency-weighted frequency over entities that previously appeared as the answer to the same (s, r) in the bank:

$$\text{HCP}_n = \alpha_{\text{hcp}} \log \left(1 + \sum_{i \in \mathcal{H}: s_i=s, r_i=r, o_i=n} e^{-\lambda \Delta_i / L} \right), \quad (13)$$

for each candidate $n \in \mathcal{E}$, with α_{hcp} a learnable scaling factor. **MES** (Memory–Entity Scoring) is an implicit dense bias that projects the retrieved memory context $\mathbf{m}$ and scores it against all entity embeddings:

$$\text{MES} = \alpha_{\text{mes}} \frac{\mathbf{Z} \mathbf{W}_m \mathbf{m}}{\sqrt{d}} \in \mathbb{R}^N, \quad (14)$$

SPS (Structural Proximity Score) bridges DHPE’s structural knowledge into scoring by measuring prototype-membership similarity between the query subject and every candidate:

$$\text{SPS} = \alpha_{\text{sps}} \cdot \mathbf{H}_e[s] \mathbf{H}_e^\top, \quad (15)$$

where $\mathbf{H}_e[s] \in \mathbb{R}^{K_e}$ is the prototype-membership vector of subject s from DHPE (detached to prevent gradient instability). Entities sharing prototype memberships with the query subject thus receive a structural prior boost.

The three signals are combined through structure-aware expert gating, where HCP and SPS are activated by query-conditioned scalar gates and MES is always on as a dense fallback:

$$\mathbf{b} = g_{\text{hcp}} \cdot \text{HCP} + \text{MES} + g_{\text{sps}} \cdot \text{SPS}, \quad (16)$$

$$\begin{aligned} g_{\text{hcp}} &= \sigma(\mathbf{w}_{\text{hcp}}^\top [\mathbf{q}; \mathbf{m}]), \\ g_{\text{sps}} &= \sigma(\mathbf{w}_{\text{sps}}^\top [\mathbf{q}; \mathbf{m}]). \end{aligned} \quad (17)$$

Both g_{hcp} and g_{sps} are initialized with a negative bias so they default to near-zero. This lets the model *adaptively* fall back on the explicit historical prior and structural prior only when the query context indicates they are useful, while MES provides a smooth dense bias for queries with no strong recurring or structural cue.

3.5 Training Objective

SHAPE is trained jointly with the backbone end-to-end. The total loss augments the backbone’s task loss with two prototype regularization terms applied to both DHPE views:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \sum_{v \in \{e, r\}} \left[\lambda_{\text{div}} \left\| \hat{\mathbf{P}}_v \hat{\mathbf{P}}_v^\top - \mathbf{I} \right\|_F^2 + \lambda_{\text{bal}} \text{Var}(\mathbf{H}_v^\top \mathbf{1}) \right]. \quad (18)$$

where $v = e$ refers to the entity-side prototype matrix $\mathbf{P}_e$ and incidence $\mathbf{H}_e$, and $v = r$ to their relation-side counterparts $\mathbf{P}_r$ and $\mathbf{H}_r$. $\hat{\mathbf{P}}_v$ denotes the row-normalized prototype matrix. The diversity term encourages prototype orthogonality, and the balance term prevents mode collapse by penalizing load imbalance across prototypes ($\mathbf{H}_v^\top \mathbf{1}$ is the per-prototype load vector). No pre-training or multi-stage schedule is required.

4 Experiments

4.1 Experiment Settings

4.1.1 Datasets

We evaluate SHAPE on three widely used TKG benchmarks: ICEWS14, ICEWS18, and ICEWS05-15. All three datasets are derived from the Integrated Crisis Early Warning System (Boschee et al., 2015) and capture geopolitical interactions over one year, one year, and eleven years, respectively. Following standard chronological splits used in prior TKG work (Li et al., 2021, 2022; Chen et al., 2024), we divide each dataset into training, validation, and test subsets with an 80%/10%/10% ratio. The details of datasets statistics are shown in Appendix A.

4.1.2 Evaluation Metrics

We report the standard time-aware filtered metrics for TKG extrapolation: Mean Reciprocal Rank (MRR) and Hits@{1,3,10}. MRR captures overall ranking quality, while Hits@ k measures whether the correct entity appears among the top- k predictions. For fair comparison, all results follow the same time-aware filtering protocol adopted in prior work (Li et al., 2021, 2022; Chen et al., 2024).

4.1.3 Baselines

We compare SHAPE with a broad set of representative KG and TKG reasoning methods, spanning static graph methods, recurrent and history-based

Model	ICEWS14				ICEWS18				ICEWS05-15			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
DistMult	27.67	18.16	31.15	46.96	10.17	4.52	10.33	21.25	28.73	19.33	32.19	47.54
ComplEx	30.84	21.51	34.48	49.59	21.01	11.87	23.47	39.97	31.69	21.44	35.74	52.04
GCN	32.15	23.07	35.53	50.63	24.99	16.04	27.86	42.75	33.67	24.08	36.99	53.67
R-GCN	28.03	19.42	31.95	44.83	15.05	8.13	16.49	29.00	27.13	18.83	30.41	43.16
ConvTransE	31.50	22.46	34.98	50.03	23.22	14.26	26.13	41.34	30.28	20.79	33.80	49.95
RE-NET	39.86	30.11	44.02	58.21	29.78	19.73	32.55	48.46	43.67	33.55	48.83	62.72
CyGNet	37.65	27.43	42.63	57.90	27.12	17.21	30.97	46.85	40.42	29.44	46.06	61.60
RE-GCN	42.00	31.63	47.20	61.65	32.62	22.39	36.79	52.68	48.03	37.33	53.90	68.51
xERTE	40.79	32.70	45.67	57.30	29.31	21.03	33.51	46.48	46.62	37.84	52.31	63.92
TiRGN	44.04	33.83	48.95	63.84	33.66	23.19	37.99	54.22	50.04	39.25	56.13	70.71
HGLS	47.00	35.06	—	70.41	29.32	19.21	—	49.83	46.21	35.32	—	67.12
RPC	44.55	34.87	49.80	65.08	34.91	24.34	38.74	54.22	51.14	39.47	57.11	71.75
DHE-TKG	40.02	30.13	44.99	—	29.23	19.73	33.31	—	45.05	34.16	50.97	—
LogCL	48.87	37.76	54.71	70.26	35.67	24.53	40.32	57.74	57.04	46.07	63.72	77.87
MGESL	45.88	35.43	51.54	65.70	34.18	23.66	38.64	54.89	53.78	42.52	60.40	75.04
CognTKE	46.06	36.49	51.11	64.49	35.24	25.21	39.93	54.71	53.13	42.62	59.42	72.70
RE-SEGNN	45.81	35.79	51.01	64.66	34.94	24.57	39.44	55.08	51.54	41.05	57.62	71.19
DiMNet	45.72	34.41	51.37	67.99	34.13	23.29	38.42	55.80	<u>58.93</u>	48.55	<u>65.16</u>	<u>78.33</u>
HGLS-SHAPE (Ours)	<u>51.08</u>	<u>40.10</u>	<u>56.67</u>	72.41	<u>37.15</u>	<u>26.32</u>	<u>41.73</u>	<u>58.64</u>	49.96	40.37	54.91	68.04
LogCL-SHAPE (Ours)	51.23	40.63	56.87	<u>72.02</u>	37.89	26.83	42.43	59.90	58.95	<u>48.43</u>	65.27	78.72

Table 1: Overall performance comparison on three ICEWS benchmarks. The best result in each column is in **bold** and the second best is underlined. All metrics are reported as percentages (%); – indicates that the metric is not reported in the original paper.

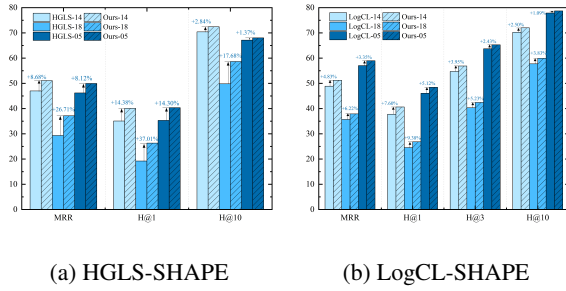


Figure 5: Relative improvements of SHAPE over two TKGC backbones on three benchmarks.

models, rule/path-based reasoning models, long-range temporal graph models, and recent methods with higher-order structure or richer temporal semantics. Among them, HGLS and LogCL are also used as the two backbone models for evaluating the plug-and-play effect of SHAPE. A detailed description of all baselines is provided in Appendix B.

4.2 Main Results

To demonstrate the effectiveness and flexibility of SHAPE, we integrated it into two leading TKGC models. Table 1 reports the overall comparison with 18 baselines across three ICEWS benchmarks, while Figure 5(a) and Figure 5(b) show the relative improvements of SHAPE over the HGLS and LogCL backbones, respectively.

As demonstrated in Table 1, SHAPE consistently improves both backbones and achieves the strongest overall performance among the compared methods.

(1) **SHAPE achieves the best MRR on all three benchmarks.** LogCL-SHAPE attains the highest MRR on ICEWS14, ICEWS18, and ICEWS05-15. Overall, the two SHAPE variants achieve the best result in 11 of the 12 columns, with the only exception being Hit@1 on ICEWS05-15, where DiMNet remains slightly higher.

(2) **SHAPE improves both backbones consistently.** Figures 5a and 5b show positive gains on every reported metric for both HGLS and LogCL. The improvements are especially pronounced on Hit@1. For example, HGLS-SHAPE improves Hit@1 by 14.38% on ICEWS14 and 37.01% on ICEWS18, while LogCL-SHAPE improves the same metric by 7.60% and 9.38%, respectively. This suggests that SHAPE does not merely reshuffle lower-ranked candidates; it more often moves the correct entity to the top of the ranking.

(3) **The gains are larger when the backbone lacks one of SHAPE’s capabilities.** HGLS benefits more strongly than LogCL, particularly on ICEWS18, where its baseline MRR is only 29.32. After adding SHAPE, HGLS rises to 37.15, corresponding to a 26.71% relative improvement. This is consistent with the design of SHAPE: DHPE contributes higher-order structural priors before temporal encoding, while QDEM improves query-conditioned history selection after encoding. Backbones that lack one or both capabilities therefore gain more from the augmentation.

(4) **The improvements generalize across differ-**

Variant	ICEWS14				ICEWS18				ICEWS05-15				$\overline{\Delta\text{MRR}}$
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	
Full SHAPE (A0)	51.08	40.10	56.67	72.41	37.15	26.32	41.73	58.64	49.96	40.37	54.91	68.04	–
w/o DHPE (A1)	47.44	36.52	52.73	68.78	35.04	24.53	39.06	56.13	45.59	36.21	49.51	63.65	–3.37
w/o QDEM (A2)	47.77	36.33	53.60	69.94	33.73	22.84	38.02	55.40	46.88	35.44	50.39	64.68	–3.27
w/o Cand. Bias (A3)	50.01	38.86	56.17	71.42	36.59	25.64	41.16	58.39	47.05	37.52	51.57	65.47	–1.52
w/o Subj. Enh. (A4)	50.60	39.21	56.78	72.43	36.12	25.25	40.64	57.44	47.48	38.03	51.94	65.59	–1.32
w/o Cross-View (A5)	50.31	39.40	55.81	71.30	36.62	25.80	41.24	58.03	47.02	37.56	51.46	65.26	–1.40
w/o Ent. Hyper. (A6)	50.71	39.42	56.59	72.41	36.13	25.35	40.51	57.45	46.88	37.39	51.48	64.81	–1.46
Backbone Only (A7)	47.00	35.06	–	70.41	29.32	19.21	–	49.83	46.21	35.32	–	67.12	–5.55

Table 2: Ablation study on the **HGLS** backbone. Best results are in **bold**. ΔMRR denotes the absolute MRR drop from Full SHAPE, averaged over three datasets.

Variant	ICEWS14				ICEWS18				ICEWS05-15				$\overline{\Delta\text{MRR}}$
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	
Full SHAPE (A0)	51.23	40.63	56.87	72.02	37.89	26.83	42.43	59.90	58.95	48.43	65.27	78.72	–
w/o DHPE (A1)	46.55	37.15	54.72	70.58	37.16	26.21	41.63	58.88	56.81	45.81	63.36	77.70	–2.52
w/o QDEM (A2)	48.55	37.23	54.54	70.23	36.27	25.07	40.94	58.47	56.73	45.63	63.54	77.74	–2.17
w/o Cand. Bias (A3)	48.83	37.51	54.53	70.87	36.16	24.84	40.77	58.64	57.11	46.01	63.92	78.21	–1.99
w/o Subj. Enh. (A4)	48.76	37.57	54.70	70.22	37.59	26.45	42.24	59.75	58.56	48.13	64.82	78.41	–1.05
w/o Cross-View (A5)	48.68	37.57	54.58	70.54	37.61	26.42	42.34	59.76	57.56	46.32	64.10	78.31	–1.41
w/o Ent. Hyper. (A6)	48.73	37.15	55.15	70.75	37.36	25.25	43.01	61.12	57.02	46.13	63.69	77.99	–1.67
Backbone Only (A7)	48.87	37.76	54.71	70.26	35.67	24.53	40.32	57.74	57.04	46.07	63.72	77.87	–2.16

Table 3: Ablation study on the **LogCL** backbone. Best results are in **bold**. ΔMRR denotes the absolute MRR drop from Full SHAPE, averaged over three datasets.

ent temporal regimes. Although the three ICEWS datasets differ in timespan, entity cardinality, and event density, SHAPE remains beneficial on all of them. This suggests that the learned prototype structure and query-conditioned event memory capture reusable reasoning biases. HGLS-SHAPE still lags behind LogCL on ICEWS05-15, but it improves HGLS by 3.75 MRR, consistent with SHAPE’s plug-and-play goal.

4.3 Ablation Study

To assess the contribution of each component in SHAPE, we benchmark seven variants on both backbones; the corresponding numbers are reported in Tables 2 and 3. Starting from the full model A0, A1 and A2 remove DHPE and QDEM, respectively. A3 and A4 disable the candidate-bias pathway and the subject-enhancement pathway in QDEM. A5 and A6 remove the cross-view semantic bridge and the entity hypergraph branch in DHPE. A7 is the bare backbone. We further conduct fine-grained ablations on LogCL-SHAPE to isolate the three candidate-scoring signals in QDEM, as reported in Table 4.

Removing either DHPE or QDEM consistently degrades MRR on both backbones, showing that the two modules are both necessary. DHPE brings a

slightly larger average gain overall, while QDEM is particularly helpful on ICEWS18, suggesting that the two modules capture complementary structural and temporal signals. The full model also outperforms the single-module variants by a clear margin, especially on LogCL, indicating that DHPE and QDEM are not redundant but mutually reinforcing: DHPE provides structural priors, and QDEM converts them into query-conditioned scoring gains.

Inside QDEM, disabling the candidate-bias pathway (A3) hurts performance more than disabling subject enhancement (A4), implying that the HCP+MES+SPS bias is the dominant contributor, while subject enhancement plays a supporting role. Inside DHPE, removing the cross-view semantic bridge (A5) and removing the entity hypergraph branch (A6) lead to comparable drops, suggesting that DHPE depends on the interaction between these two parts rather than on either one alone.

Table 4 isolates the three candidate-scoring signals on LogCL-SHAPE. Removing HCP, MES, or SPS reduces the average MRR by 0.72, 0.92, and 0.71 percentage points, respectively. Full SHAPE achieves the highest MRR and Hits@1, indicating that the three signals provide complementary evidence. HistFreq-only improves over the LogCL backbone by 0.63, 1.01, and 0.43 MRR points on

Variant	ICEWS14		ICEWS18		ICEWS05-15	
	MRR	H@1	MRR	H@1	MRR	H@1
Full SHAPE	51.23	40.63	37.89	26.83	58.95	48.43
w/o HCP	50.01	39.45	37.43	26.06	58.47	47.99
w/o MES	49.84	38.61	37.30	26.29	58.16	47.42
w/o SPS	49.88	39.12	37.60	26.51	58.45	47.06
HistFreq-only	49.50	39.05	36.68	25.33	57.47	46.56

Table 4: Fine-grained ablation of the QDEM candidate-scoring signals on LogCL-SHAPE. HistFreq-only retains only the recency-weighted historical-frequency bias. The best results are shown in **bold**.

ICEWS14, ICEWS18, and ICEWS05-15, respectively. However, it remains 1.73, 1.21, and 1.48 points below Full SHAPE. Historical repetition is useful but cannot fully account for SHAPE’s gains. Memory–entity matching and structural prototype proximity provide additional improvements.

5 Conclusion

We present SHAPE, a plug-and-play enhancement framework for TKG reasoning that combines higher-order structural augmentation with query-conditioned event denoising. In particular, DHPE learns prototype hypergraphs for entities and relations, and injects cross-view structural priors before temporal encoding to enrich representation learning. QDEM further retrieves query-relevant recent events and transforms them into subject enhancement and candidate-level bias after encoding, thereby refining predictions with query-specific temporal evidence. Extensive experiments on three datasets demonstrate the effectiveness of our model.

Limitations

SHAPE is designed as a plug-and-play enhancement framework, so its effectiveness naturally depends on the quality of the underlying backbone. When the backbone already captures strong structural or temporal dependencies, the marginal gains brought by SHAPE may be limited. Therefore, in our future work, we will focus on how to better adapt SHAPE to stronger backbone models.

Acknowledgments

This study was supported in part by the National Science Fund for Distinguished Young Scholars (62325308), in part by the State Key Program of the National Natural Science Foundation of China (62432013), and in part by the National Natural Science Foundation of China (62522316, 62472353, 62572399).

References

- Elizabeth Boschee, Jennifer Lautenschlager, Sean O’Brien, Steve Shellman, James Starz, and Michael Ward. 2015. [ICEWS coded event data](#).
- Wenyu Cai, Mengfan Li, Xuanhua Shi, Yuanxin Fan, Quntao Zhu, and Hai Jin. 2025. [RE-SEGNN: recurrent semantic evidence-aware graph neural network for temporal knowledge graph forecasting](#). *Sci. China Inf. Sci.*, 68(2).
- Wei Chen, Huaiyu Wan, Yuting Wu, Shuyuan Zhao, Jiayao Cheng, Yuxin Li, and Youfang Lin. 2024. [Local-global history-aware contrastive learning for temporal knowledge graph reasoning](#). In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*, pages 733–746. IEEE.
- Wei Chen, Yuting Wu, Shuhan Wu, Zhiyu Zhang, Mengqi Liao, Youfang Lin, and Huaiyu Wan. 2025. [CognTKE: A cognitive temporal knowledge extrapolation framework](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 14815–14823. AAAI Press.
- Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. 2022. You are AllSet: A multiset function framework for hypergraph neural networks. In *Proceedings of the 10th International Conference on Learning Representations*.
- Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. [HyTE: Hyperplane-based temporally aware knowledge graph embedding](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2001–2011, Brussels, Belgium. Association for Computational Linguistics.
- Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2020. [Dynamic knowledge graph based multi-event forecasting](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’20*, page 1585–1595, New York, NY, USA. Association for Computing Machinery.
- Hao Dong, Ziyue Qiao, Zhiyuan Ning, Qi Hao, Yi Du, Pengyang Wang, and Yuanchun Zhou. 2025. [Dis-entangled multi-span evolutionary network against temporal knowledge graph reasoning](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 12697–12707. Association for Computational Linguistics.
- Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. [Hypergraph neural networks](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative*

- Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.
- Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. 2021. [Explainable subgraph reasoning for forecasting on temporal knowledge graphs](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Xingyue Huang, Mikhail Galkin, Michael M. Bronstein, and Ismail Ilkan Ceylan. 2026. [HYPER: A foundation model for inductive link prediction with knowledge hypergraphs](#). In *The Fourteenth International Conference on Learning Representations*.
- Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. 2020. [Temporal Knowledge Base Completion: New Algorithms and Evaluation Protocols](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3733–3747, Online. Association for Computational Linguistics.
- Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. 2020. [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, Online. Association for Computational Linguistics.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *International Conference on Learning Representations*.
- Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. 2020. [Tensor decompositions for temporal knowledge base completion](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Yujia Li, Shiliang Sun, and Jing Zhao. 2022. [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. ijcai.org.
- Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 2021. [Temporal knowledge graph reasoning based on evolutionary 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. 2023. [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 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1559–1568. ACM.
- Xinyue Liu, Jianan Zhang, Chi Ma, Wenxin Liang, Bo Xu, and Linlin Zong. 2024. [Temporal knowledge graph reasoning with dynamic hypergraph embedding](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15742–15751. ELRA and ICCL.
- Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. 2022. [TLogic: temporal logical rules for explainable link forecasting on temporal knowledge graphs](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 4120–4127. AAAI Press.
- Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling relational data with graph convolutional networks](#). In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings*, page 593–607, Berlin, Heidelberg. Springer-Verlag.
- Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. 2019. [End-to-end structure-aware convolutional networks for knowledge base completion](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19*. AAAI Press.
- Mingcong Shi, Chunjiang Zhu, Detian Zhang, Shiting Wen, and Qing Li. 2024. [Multi-granularity history and entity similarity learning for temporal knowledge graph reasoning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5232–5243. Association for Computational Linguistics.
- 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8306–8319. Association for Computational Linguistics.
- Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, Liming Zhu, and Wenjie Zhang. 2026.

- MemoTime: Memory-augmented temporal knowledge graph enhanced large language model reasoning.** In *Proceedings of the ACM Web Conference 2026*, WWW '26, pages 4220–4231, New York, NY, USA. Association for Computing Machinery.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2071–2080. JMLR.org.
- Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. **Kgat: Knowledge graph attention network for recommendation.** In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 950–958, New York, NY, USA. Association for Computing Machinery.
- Chengjin Xu, Yung-Yu Chen, Mojtaba Nayyeri, and Jens Lehmann. 2021. **Temporal knowledge graph completion using a linear temporal regularizer and multivector embeddings.** In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2569–2578, Online. Association for Computational Linguistics.
- Yi Xu, Junjie Ou, Hui Xu, and Luoyi Fu. 2023. **Temporal knowledge graph reasoning with historical contrastive learning.** In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 4765–4773. AAAI Press.
- Naganand Yadati, Tarun Kumar, Deepak Maurya, Balaraman Ravindran, and Partha Talukdar. 2024. **HEAL: unlocking the potential of learning on hypergraphs enriched with attributes and layers.** In *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pages 34:1–34:25. PMLR.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. **Embedding Entities and Relations for Learning and Inference in Knowledge Bases.** In *International Conference on Learning Representations*.
- 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)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 12617–12631. 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 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023, pages 2412–2422. ACM.
- Cunchao Zhu, Muhao Chen, Changjun Fan, Guangquan Cheng, and Yan Zhang. 2021. **Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks.** In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4732–4740. AAAI Press.

A Datasets

The details of the dataset used in the experiment are shown in Table 5.

Datasets	Entities	Relations	Total Facts	Train Facts	Time Steps
ICEWS14	7128	230	90730	74845	365
ICEWS05-15	10488	251	479329	368868	4017
ICEWS18	23033	256	468558	373018	304

Table 5: Statistics of the benchmark datasets.

B Baselines

We compare SHAPE with a broad set of representative KG and TKG reasoning methods. The static and graph-based baselines include DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), GCN (Kipf and Welling, 2017), R-GCN (Schlichtkrull et al., 2018), and ConvTransE (Shang et al., 2019). These methods cover widely used embedding, convolutional, and relation-aware graph encoders, and therefore provide reference points for evaluating how much temporal modeling and structural augmentation improve over non-temporal or weakly temporal scoring functions.

For TKG extrapolation, we include methods that represent several major modeling families: recurrent or history-vocabulary models (RE-NET (Jin et al., 2020), CyGNet (Zhu et al., 2021), RE-GCN (Li et al., 2021), and TiRGN (Li et al., 2022)), query-aware or rule/path-based reasoning models (xERTE (Han et al., 2021) and RPC (Liang et al., 2023)), long-range temporal graph models (HGLS (Zhang et al., 2023b) and LogCL (Chen et al., 2024)), and recent models with higher-order structure or richer temporal semantics (DHE-TKG (Liu et al., 2024), MGESL (Shi et al., 2024), CognTKE (Chen et al., 2025), RE-SEGNN (Cai et al., 2025), and DiMNet (Dong et al., 2025)).

For baselines without results on the considered ICEWS benchmarks in their original papers, we

use results reported by prior published TKG studies under the time-aware filtered protocol. Otherwise, results are taken from the corresponding original papers. Our main comparisons between HGLS and HGLS-SHAPE, as well as between LogCL and LogCL-SHAPE, are conducted with identical backbone implementations, data splits, and evaluation code within each pair.

C Implementation Details

For both backbones, the hidden dimension is fixed to $d = 200$. We use dataset-specific SHAPE configurations on the three ICEWS benchmarks. For the optimal results reported in the paper, the settings are as follows. On ICEWS14, we set $K_e = 256$, $K_r = 32$, $h = 4$, and $L = 3$. On ICEWS18, we set $K_e = 256$, $K_r = 32$, $h = 4$, and $L = 8$. On ICEWS05-15, we set $K_e = 256$, $K_r = 64$, $h = 4$, and $L = 10$. The regularization weights are $\lambda_{\text{div}} = \lambda_{\text{bal}} = 0.01$. All models are trained with Adam (lr = 0.001, weight decay 10^{-5}) and early stopping (patience = 10 epochs, monitoring validation MRR). Gradient norms are clipped at 1.0.

D Evaluation on an Additional Backbone

To further assess the architectural generality of SHAPE, we integrate it into RE-GCN, a recurrent evolution-based TKG backbone, and evaluate both variants under the time-aware filtered protocol.

Dataset	Model	MRR	Hits@1	Hits@3	Hits@10
ICEWS14	RE-GCN	42.00	31.63	47.20	61.65
	RE-GCN-SHAPE	45.31	34.97	50.87	64.80
ICEWS18	RE-GCN	32.62	22.39	36.79	52.68
	RE-GCN-SHAPE	35.03	24.66	39.40	55.42
ICEWS05-15	RE-GCN	48.03	37.33	53.90	68.51
	RE-GCN-SHAPE	50.52	40.48	56.04	69.47

Table 6: Performance on the additional RE-GCN backbone under time-aware filtered evaluation.

RE-GCN-SHAPE improves all 12 metrics, with MRR gains of 3.31, 2.41, and 2.49 points on the three datasets, respectively. These results provide further evidence that SHAPE transfers effectively to a recurrent evolution-based backbone.

E Sensitivity Analysis

We study three hyperparameters that directly control the capacity of SHAPE: the number of entity prototypes K_e , the number of relation prototypes K_r , and the event memory length L . All sensitivity experiments are conducted on LogCL-SHAPE

while varying one hyperparameter at a time. Figures 6 and 7 report the results on ICEWS14, ICEWS18, and ICEWS05-15.

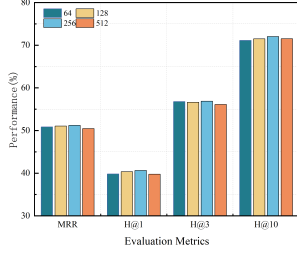
Entity prototype capacity. The top row of Figure 6 shows a consistent pattern across all three datasets: increasing K_e from 64 or 128 to 256 improves the results, while further increasing it to 512 brings no additional benefit and slightly reduces most metrics. This indicates that DHPE needs enough prototypes to represent recurring entity roles and higher-order semantic groups, but excessive entity prototypes mainly introduce redundant partitions of an already learnable structure. The stability of the optimum across ICEWS14, ICEWS18, and ICEWS05-15 is important: it suggests that $K_e = 256$ is not a dataset-specific accident, but a robust capacity level for ICEWS-style event graphs with thousands to tens of thousands of entities.

Relation prototype capacity. The bottom row of Figure 6 shows that relation prototypes require a smaller capacity than entity prototypes. ICEWS14 and ICEWS18 achieve the best performance at $K_r = 32$, while ICEWS05-15 favors $K_r = 64$. This trend is consistent with the structure of temporal event graphs, where the relation vocabulary is much smaller and semantically more regular than the entity vocabulary, making a compact set of relation prototypes sufficient for most datasets. The larger optimum on ICEWS05-15 is also reasonable, since this dataset spans a much longer temporal period and exhibits more diverse relation usage patterns over time, which increases the demand for prototype diversity. Overall, these results support the asymmetric design of DHPE, in which entity and relation hypergraphs are both necessary, but their prototype capacities should not be set to the same value.

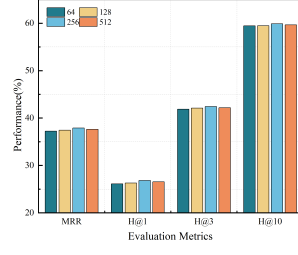
Model	#Params	#SHAPE	Epochs	Time/Ep.	Total Time	Mem
HGLS	5.97M	–	32	99.9s	76.5min	2.18GB
HGLS-SHAPE	8.65M	2.72M	33	99.3s	78.2min	2.37GB
LogCL	8.28M	–	22	229.5s	90.2min	1.54GB
LogCL-SHAPE	11.00M	2.72M	22	234.2s	91.2min	1.95GB

Table 7: Efficiency comparison on ICEWS14.

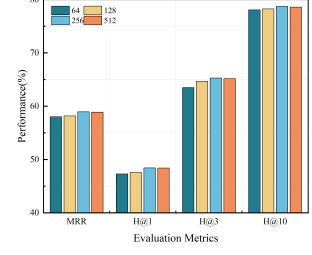
Event memory length. Figure 7 shows that the preferred memory length is dataset-dependent, but the overall behavior is stable. ICEWS14 favors a short window around $L = 3$, ICEWS18 benefits from a medium window around $L = 8$, and ICEWS05-15 reaches its best results around $L = 10$. These differences match the temporal scale of the benchmarks: shorter one-year event streams often concentrate useful evidence in very recent snapshots,



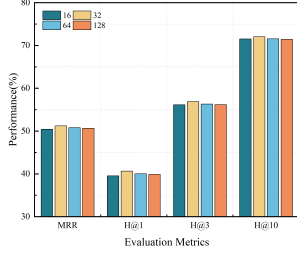
(a) ICEWS14, entity prototypes



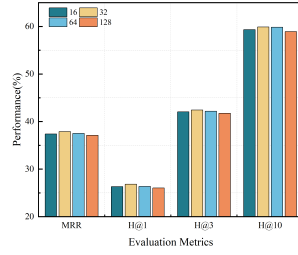
(b) ICEWS18, entity prototypes



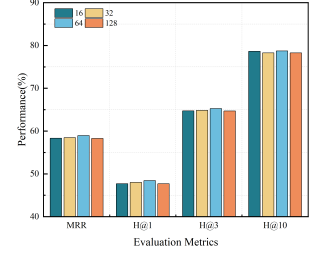
(c) ICEWS05-15, entity prototypes



(d) ICEWS14, relation prototypes



(e) ICEWS18, relation prototypes

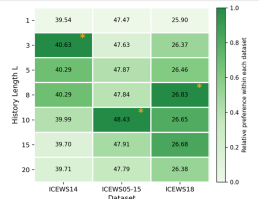


(f) ICEWS05-15, relation prototypes

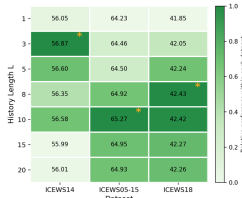
Figure 6: Sensitivity to the number of entity and relation prototypes on LogCL-SHAPE. The top row varies $K_e \in \{64, 128, 256, 512\}$, and the bottom row varies $K_r \in \{16, 32, 64, 128\}$.



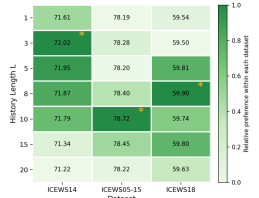
(a) MRR



(b) Hits@1



(c) Hits@3



(d) Hits@10

Figure 7: Sensitivity to the event memory length L on LogCL-SHAPE. Each heatmap is normalized within each dataset to highlight the preferred history-length region, while the cell values report the original metric scores.

whereas the longer ICEWS05-15 sequence benefits from a wider context that can cover slower recurring patterns. More importantly, performance tends to stabilize after the effective historical range is covered rather than degrading sharply as L increases. This behavior supports the design of QDEM. Since QDEM performs query-conditioned top- k retrieval instead of indiscriminately pooling all events in the window, increasing L expands the candidate memory bank but does not force the decoder to consume every historical event. As a result, SHAPE is ro-

bust to the memory length once the window is large enough to include the relevant temporal evidence.

F Robustness under Data Sparsity

To evaluate whether SHAPE’s improvements stem from genuine structural and historical priors rather than merely fitting more parameters to abundant data, we conduct a data sparsity experiment on ICEWS14. We randomly remove $\{10\%, 30\%, 50\%, 70\%, 90\%\}$ of the training triples (preserving temporal structure by sampling independently within each snapshot) and compare both backbones with and without SHAPE. The results are shown in Figure 8.

(1) SHAPE remains beneficial across all sparsity levels. Across all three metrics and both backbones, SHAPE-enhanced models consistently outperform their corresponding baselines at the original setting and under 10%–90% deletion. This makes it unlikely that the gains are driven merely by parameter count. This robustness can be attributed to the complementary nature of DHPE and QDEM under data-scarce conditions

(2) The structural prior becomes more valuable for HGLS under moderate sparsity. For HGLS, the MRR gap widens from +4.08 at the full-data setting to +6.60 at 30% deletion. A similar trend appears on Hit@1. This is consistent with the role of DHPE: when local neighborhoods become sparse,

Query	Truth	Rank _B	Rank _S	Backbone Top-3	SHAPE Top-3
s: Ali_Tayyebnia r: Express_intent_to_cooperate t: 337	Iran	192	1	Shahid Khan; MP (India); Media (Pakistan)	Iran; Business (Iran); Citizen (India)
s: Public_Prosecutor_(Thailand) r: Express_intent_to_cooperate t: 359	Citizen_(Thailand)	47	1	Employee (India); Armed rebel (NSCN); F. Fasehun	Citizen (Thailand); Thailand; Public Prosecutor (Thailand)
s: Member_of_Parliament_(Russia) r: Make_statement t: 342	RIA_Novosti	45	1	F. Mogherini; Gov. (Ukraine); Interfax	RIA Novosti; Ministry (Russia); Interfax
s: Amara_Essy r: Use_conventional_military_force t: 356	Citizen_(Cote_d'Ivoire)	38	1	Citizen (Tanzania); Citizen (India); Citizen (Botswana)	Citizen (Cote d'Ivoire); Criminal (Mexico); Citizen (Kenya)
s: Ministry_(France) r: Engage_in_symbolic_act t: 341	Francois_Hollande	27	1	Ministry (France); Military (France); Student (France)	Francois Hollande; Citizen (UK); Military (France)
s: Government_(United_States) r: Accuse t: 353	North_Korea	26	1	Bahrain; Citizen (Kenya); Human Rights Watch	North Korea; Boko Haram; China

Table 8: Case study on ICEWS14. Rank_B and Rank_S denote the filtered ranks assigned by the LogCL backbone and LogCL-SHAPE, respectively.

Analysis	Query group	MRR _B	MRR _S	Δ MRR	Δ H@1 (pp)
Overall	All object queries	0.4879	0.5123	+0.0244	+2.62
Error severity	Rank _B = 2–3	0.4450	0.5966	+0.1515	+37.86
	Rank _B = 4–10	0.1764	0.3231	+0.1467	+14.46
	Rank _B = 11–30	0.0604	0.1369	+0.0765	+3.84
	Rank _B = 31–100	0.0204	0.0540	+0.0336	+0.58
	Rank _B > 100	0.0033	0.0196	+0.0163	+0.67
Recent evidence	Exact answer observed in recent history	0.7282	0.8407	+0.1125	+15.05
Relation type	Make optimistic comment	0.3644	0.4551	+0.0907	+12.12
	Meet at a third location	0.6566	0.7397	+0.0831	+12.33
	Discuss by telephone	0.7726	0.8379	+0.0653	+9.38

Table 9: Systematic analysis of object-prediction queries on ICEWS14. MRR_B and MRR_S denote the results of LogCL and LogCL-SHAPE, respectively. Δ H@1 denotes the absolute percentage-point improvement. The relation-type rows report selected relations with substantial gains.

prototype sharing provides an additional channel for transferring structural information.

(3) QDEM helps LogCL degrade more gracefully under severe sparsity. At 90% deletion, LogCL-SHAPE retains a higher proportion of its original performance than the bare backbone. This resilience is attributable to the design of QDEM’s biased relevance scoring, which explicitly rewards events matching the query subject or relation and therefore concentrates the model’s attention on the most informative subset of available history.

G Efficiency Analysis

A natural concern is whether SHAPE’s performance gains come at prohibitive computational cost. We report the parameter count, convergence speed, total training time, and peak GPU memory on ICEWS14 for both backbones in Table 7. All experiments are conducted on a single NVIDIA RTX 4090 GPU with identical batch sizes, data loaders, and early-stopping configurations. #Params is the total trainable parameter count; #SHAPE denotes

the additional parameters introduced by DHPE and QDEM; Epochs is the number of epochs to converge; Time/Ep. is the average wall-clock training time per epoch (excluding validation); Total Time is the end-to-end training wall-clock time; Mem is the peak GPU memory.

SHAPE introduces approximately 2.72M additional trainable parameters on both backbones, comprising the DHPE prototype matrices, cross-view bridge weights, and QDEM event-memory parameters. This corresponds to 31.5% of the HGLS-SHAPE parameters and 24.7% of the LogCL-SHAPE parameters, which is a moderate increase relative to the performance gains reported in Section 4.2.

Per-epoch runtime remains comparable after adding SHAPE: the difference is below 1% on HGLS and around 2% on LogCL. The number of epochs to convergence also remains nearly unchanged. The memory increase is modest (+0.19 GB on HGLS and +0.41 GB on LogCL), with the main overhead coming from the soft inci-

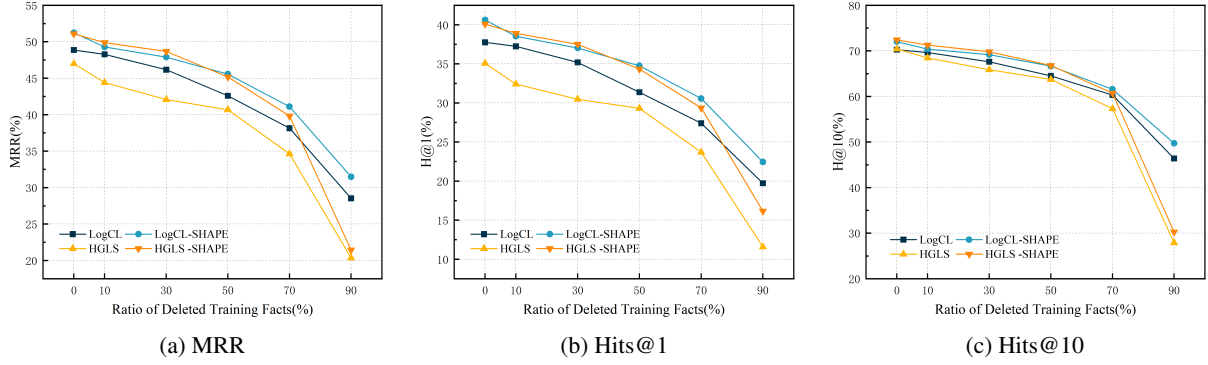


Figure 8: Performance under varying training data sparsity on ICEWS14. The x-axis denotes the ratio of randomly deleted training facts. Each subplot shows four curves: HGLS, HGLS-SHAPE, LogCL, and LogCL-SHAPE.

dence matrix $\mathbf{H}_e \in \mathbb{R}^{N \times K_e}$.

Overall, the efficiency analysis shows that SHAPE delivers consistent accuracy gains with only modest additional computation, which is important for a plug-and-play module intended for existing TKG backbones.

H Case Study

To provide qualitative evidence of how SHAPE improves predictions, we select representative queries from the ICEWS14 test set under a strict criterion: the LogCL backbone ranks the correct answer outside the top 20, while LogCL-SHAPE moves it to rank 1. Table 8 lists such examples together with the top-3 predictions of each model. Here, t denotes the timestamp identifier in the dataset. These cases are particularly informative because they highlight situations in which the backbone fails to identify the correct entity, whereas SHAPE is able to recover it reliably.

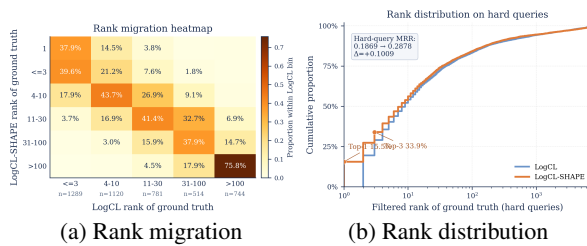


Figure 9: Case-study visualization on hard ICEWS14 queries.

The table and Figure 9 reveal two recurring patterns. First, SHAPE can correct severe ranking errors rather than merely refine already strong predictions: in all listed cases, the correct entity is ranked outside the top 20 by the backbone but is promoted to the first position after applying SHAPE. This behavior is also reflected in Figure 9(a). For example, among queries whose ground-truth entities

are ranked within the 4–10 bin by LogCL, 14.5% are promoted to rank 1 and another 21.2% are promoted to the ≤ 3 bin by LogCL-SHAPE. Similar upward migration appears in harder bins, indicating that SHAPE improves a non-trivial portion of queries that the backbone does not solve confidently.

Second, the backbone often prefers plausible but incorrect entities that are topically related to the query, whereas SHAPE more reliably retrieves the ground-truth target. Figure 9(b) further supports this observation at the distribution level: on hard queries, LogCL-SHAPE raises the Top-1 and Top-3 proportions to 15.5% and 33.9%, respectively, and improves the hard-query MRR by +0.1009. These results suggest that higher-order structural priors and query-conditioned memory denoising help the model disambiguate between superficially similar candidates. Overall, SHAPE improves not only ranking quality but also the model’s ability to identify the truly relevant entity under challenging query conditions.

Table 9 extends the case study to all object-prediction queries in ICEWS14. SHAPE increases the overall MRR from 0.4879 to 0.5123 and improves Hits@1 by 2.62 percentage points. It improves MRR across every error-severity interval; among the 4,448 queries not ranked first by LogCL, 54.72% receive a better rank and 15.47% reach rank 1. Gains are particularly large when the exact answer appears in recent history, with an MRR gain of 0.1125 and a 15.05-point Hits@1 improvement. Several diplomatic-contact relations also exhibit substantial gains. These results show that the examples in Table 8 reflect broader rank-correction patterns rather than isolated cases.