

ELITE: Evolving Localized and Informative Temporal Entity alignment

Jiayun Li¹ Wen Hua^{2*} Shiqi Fan² Yuchuang Zhao¹
Wei Jiang¹ Haiyang Jiang¹ Xue Li¹

¹The University of Queensland, Brisbane, Queensland, Australia

²The Hong Kong Polytechnic University, Hong Kong SAR, China

{uqjli48, y.c.zhao, weijiang, haiyang.jiang}@uq.edu.au
{wency.hua, shiqi.fan}@polyu.edu.hk
xueli@eecs.uq.edu.au

Abstract

Temporal Entity Alignment (TEA) aims to identify equivalent entities across Temporal Knowledge Graphs (TKGs). Existing TEA methods generally assume static snapshots and repeatedly perform full-graph alignment, resulting in high computational costs. Recent evolving EA models handle continuously changing graphs through incremental alignment strategies, but mainly focus on non-temporal KGs and overlook temporal dynamics. Consequently, Evolving TEA, a realistic scenario where both graph structures and temporal information evolve over time, remains underexplored. To bridge this gap, we propose a lightweight continual TEA framework with three novel design: (1) informative seed selection that jointly leverages structural and temporal signals for reliable iterative fine-tuning, (2) localized inference that restricts computation to newly updated entities and their neighborhoods for efficient alignment, and (3) cross-snapshot score fusion that adaptively integrates historical and current alignment evidence. Extensive experiments on real-world evolving TEA benchmarks demonstrate the superiority of our proposed framework, achieving state-of-the-art alignment performance while significantly reducing computational overhead. The code and datasets are available at <https://github.com/psych-mind/ELITE>.

1 Introduction

Temporal Knowledge Graphs (TKGs) extend conventional knowledge bases by associating facts with timestamps, capturing the dynamic interactions between real-world entities. Given TKGs constructed from diverse sources, integrating knowledge across multiple TKGs has become essential. Temporal Entity Alignment (TEA) — the task of identifying equivalent entities across TKGs — plays a central role in this integration. However, existing TEA approaches (Xu et al., 2021, 2022; Cai

et al., 2022; Liu et al., 2023; Zeng et al., 2024; Cai et al., 2023; Li et al., 2025, 2026) typically operate under a fixed snapshot assumption, contradicting with the dynamic nature of TKGs.

In practice, TKGs evolve continuously as new entities and temporal facts emerge over time. Retraining TEA models for each updated snapshot is naturally time-consuming. Pioneer works on Evolving EA have attempted to incrementally maintain alignment models across snapshots by either adjusting the adjacency matrix under a training-free paradigm (Yan et al., 2021) or fine-tuning model parameters with newly-selected alignment seeds (Wang et al., 2022). However, both approaches are designed exclusively for non-temporal KGs and simulate knowledge evolution by randomly appending triples without temporal ordering, thus failing to reflect the chronological updates of real-world TKGs. A more recent effort (Li et al., 2023) targets evolving TKGs directly via a teacher-student distillation scheme, but its experiments only consider a single increment from one base snapshot to one future snapshot, rather than a continual stream of chronologically ordered updates; such a two-snapshot setting cannot capture the compounding, ongoing evolution that characterizes real-world TKGs. This calls for a more realistic yet more challenging setting of *Evolving TEA*.

Issue 1: How to select informative seeds for model fine-tuning. Overconfident seeds reinforce current predictions and hurt generalization, while low-confidence seeds introduce error propagation that gradually deteriorates alignment quality over time. This trade-off is a fundamental challenge for evolving TEA. Moreover, structural updates and temporal evolution of TKGs offer complementary signals: the former captures changes in graph topology and local semantic context, while the latter reflects consistency and dependency across time. Therefore, an effective seed selection strategy for evolving TEA should consider both structural and

*Corresponding author.

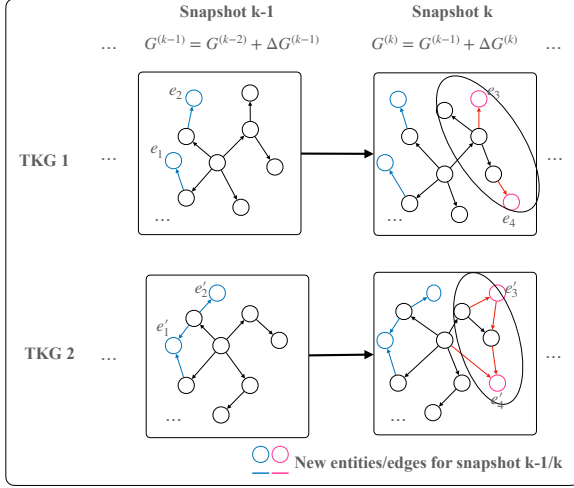


Figure 1: Illustration of two TKGs evolving across consecutive snapshots, where $\mathcal{G}^{(k)} = \mathcal{G}^{(k-1)} + \Delta\mathcal{G}^{(k)}$ incrementally incorporates new entities and temporal facts (blue: snapshot $k-1$; red: snapshot k). Existing entity pairs $e_1 \leftrightarrow e'_1$ and $e_2 \leftrightarrow e'_2$ are already aligned in snapshot $k-1$, while e_3, e_4 and e'_3, e'_4 emerge as new unaligned entities in snapshot k . Our goal is to refine the alignment of existing pairs while establishing alignment for newly emerging ones.

temporal evidence, so as to identify reliable yet informative seeds that support robust model adaptation without amplifying prediction errors.

Issue 2: How to perform efficient localized inference. As snapshots accumulate, re-aligning all entities from scratch becomes prohibitively expensive, making full-graph inference impractical for real-time alignment. As evidenced by Figure 1 and Table 9 (Appendix D), TKG updates at each snapshot are predominantly concentrated in a small evolving region, while most existing alignment predictions remain unchanged. This motivates a localized inference strategy that focuses computation only on affected entities and their neighborhoods, reducing unnecessary processing while improving both alignment efficiency and effectiveness.

Issue 3: How to guarantee alignment quality over time. Reliable aggregation across snapshots is crucial in evolving TEA, since alignment decisions should reflect both current observations and accumulated historical evidence. Otherwise, the model becomes vulnerable to unstable updates from newly arrived data. In particular, naive aggregation that treats all snapshot-level predictions equally may mix high-confidence historical alignments with unreliable incoming scores, allowing noisy updates to overwrite trustworthy evidence and gradually degrade alignment quality. Therefore, an effective aggregation strategy is necessary

to selectively preserve reliable past evidence while suppressing low-confidence updates over time.

We propose **ELITE**, an **E**volving **L**ocalized and **I**nformative **T**emporal **E**ntity alignment framework to address these issues. In particular, we devise a *dual-aspect complementary seed selection* approach that jointly exploits relational and temporal signals to select informative and diverse alignment seeds for model fine-tuning. We introduce a *localized inference* strategy that restricts computation to selected evolving entities and their local neighborhoods, substantially reducing inference time and memory overhead on large-scale TKGs. We also propose a *cross-snapshot inference aggregation* mechanism that accumulates historical evidence while selectively suppressing low-confidence updates over time, maintaining alignment quality.

Contributions:

- We are the first to formalize and explore *Evolving TEA* under realistic, continual multi-snapshot evolution, identifying its key challenges and the limitations of directly applying existing solutions.
- We propose ELITE, an iterative TEA framework that maintains the alignment model incrementally. The novel designs of dual-aspect informative seed selection and localized cross-snapshot inference guarantees the effectiveness and efficiency of continuous alignment.
- We construct two evolving TEA benchmarks with chronological updates of real-world TKGs, providing a realistic evaluation platform for future research. Experimental results demonstrate that ELITE achieves state-of-the-art alignment accuracy while reducing inference time by $3.7\times$ and memory cost by $10\times$.

2 Problem Formulation

Definition 1 (Temporal Knowledge Graph). A *temporal knowledge graph (TKG)* is a directed graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{F})$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ denote the set of entities, relations, and timestamps, respectively, and $\mathcal{F} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathcal{T}$ is the set of temporal facts, where each fact $f = (e_h, r, e_t, t)$ records the valid relation r between head entity e_h and tail entity e_t during the time t .

Definition 2 (Evolving Temporal Knowledge Graph). An *evolving temporal knowledge graph (ETKG)* is a sequence of TKG snapshots $\{\mathcal{G}^{(k)}\}_{k=1}^n$ that grow chronologically with k , i.e., $t^{(1)} < \dots < t^{(n)}$. Each snapshot evolves as $\mathcal{G}^{(k)} = \mathcal{G}^{(k-1)} \cup$

$\Delta\mathcal{G}^{(k)}$, where $\Delta\mathcal{G}^{(k)} = (\Delta\mathcal{E}^{(k)}, \mathcal{R}, \Delta\mathcal{F}^{(k)})$ captures newly emerged entities $\Delta\mathcal{E}^{(k)} = \mathcal{E}^{(k)} \setminus \mathcal{E}^{(k-1)}$ and newly occurred temporal facts $\Delta\mathcal{F}^{(k)} = \mathcal{F}^{(k)} \setminus \mathcal{F}^{(k-1)}$ within $(t^{(k-1)}, t^{(k)}]$.

Definition 3 (Evolving Temporal Entity Alignment). *Given two ETKGs with aligned snapshot sequences $\{\mathcal{G}^{(k)}\}_{k=1}^n$ and $\{\mathcal{G}'^{(k)}\}_{k=1}^n$, the task of Evolving TEA aims to continuously identify the matchable entity pairs for each snapshot: $\{(e, e') \mid e \equiv e', e \in \mathcal{E}^{(k)}, e' \in \mathcal{E}'^{(k)}\}$.*

3 The ELITE Model

3.1 Overall Framework

State-of-the-art TEA models typically generate two complementary alignment views: a relational similarity matrix $S_r \in \mathbf{R}^{|\mathcal{E}| \times |\mathcal{E}|}$ and a temporal similarity matrix $S_t \in \mathbf{R}^{|\mathcal{E}| \times |\mathcal{E}|}$ (Liu et al., 2023; Li et al., 2025, 2026), which are subsequently integrated to produce the final alignment dictionary $\hat{D}$ that stores the top-ranked counterparts for each entity. A straightforward extension to ETKGs requires to fully retrain the model and realign the entire graph at every snapshot. However, such a strategy is computationally prohibitive and impractical for large-scale ETKGs.

As illustrated in Figure 2, ELITE iteratively fine-tunes the backbone TEA model with newly selected pseudo seeds for any two consecutive snapshots $\mathcal{G}^{k-1}$ and $\mathcal{G}^k$, and updates alignment incrementally over the evolving region while preserving evidence accumulated from previous snapshots. In particular, ELITE comprises three key modules to guarantee the alignment quality and inference efficiency. Given a snapshot $\mathcal{G}^{(k)} = \mathcal{G}^{(k-1)} \cup \Delta\mathcal{G}^{(k)}$, it first selects **informative and complementary seeds** based on the relational and temporal similarity matrices S_r^{k-1} and S_t^{k-1} , to support localized fine-tuning of the alignment model. It then performs **localized inference** with the updated model to generate localized alignment results D_k , restricting computation to evolving entities and their local neighborhoods to substantially reduce inference cost. Finally, the **cross-snapshot inference aggregation** integrates cumulative and per-snapshot scores via confidence margin-based conflict resolution, yielding globally-coherent alignment $\hat{D}_k$.

3.2 Informative Seed Selection for Fine-tuning

As the graph evolves over time, each new snapshot introduces additional triples reflecting the structural and temporal changes of that period. To ensure the

Algorithm 1: Informative seed selection

Input: S_r, S_t : relational and temporal similarity matrices of entities in the localized area; n : number of seeds to select
Output: $\hat{P}_r^{(n)}, \hat{P}_t^{(n)}$: top- n relational and temporal seed pairs; $\mathcal{R}_r, \mathcal{R}_t$: ranked seed lists

```

1  $\hat{P}_r \leftarrow []; \hat{P}_t \leftarrow [];$ 
2 for  $e \in \mathcal{E}$  do
    // choose bi-directional confident seeds per aspect
3    $e'_r \leftarrow \text{BiDir}(S_r, e);$ 
4    $e'_t \leftarrow \text{BiDir}(S_t, e);$ 
    // filter out dual-aspect confident seeds
5   if  $e = e'_r = e'_t$  then continue;
6   if  $e = e'_r$  then
7      $\hat{P}_r \leftarrow \hat{P}_r \cup \{(e, e'_r)\}$ 
8   if  $e = e'_t$  then
9      $\hat{P}_t \leftarrow \hat{P}_t \cup \{(e, e'_t)\}$ 
    // ranking the entity pairs based on complementary score
10  $\mathcal{R}_r \leftarrow \text{sort}(\{(e, e') \mapsto \text{score}(S_t(e, e'), S_r(e, e')) \mid (e, e') \in \hat{P}_r\}, \text{desc});$ 
11  $\mathcal{R}_t \leftarrow \text{sort}(\{(e, e') \mapsto \text{score}(S_t(e, e'), S_r(e, e')) \mid (e, e') \in \hat{P}_t\}, \text{desc});$ 
12  $\hat{P}_r^{(n)} \leftarrow \mathcal{R}_r[1:n]; \hat{P}_t^{(n)} \leftarrow \mathcal{R}_t[1:n];$ 
13 return  $\hat{P}_r^{(n)}, \hat{P}_t^{(n)}, \mathcal{R}_r, \mathcal{R}_t;$ 
```

model effectively incorporates this new information, we propose a fine-tuning strategy tailored to each snapshot with newly-selected pseudo seeds.

Inspired by localized graph augmentation (Liu et al., 2022) and active learning (Settles, 2009; Liu et al., 2021), we propose a *localized informative seed selection* strategy for snapshot fine-tuning. Rather than selecting the most confident pairs, we argue that seeds at the boundary of model certainty carry richer information about the evolving snapshot. However, purely uncertain seeds risk noisy supervision and error propagation. To balance quality and diversity, we exploit the complementary relational and temporal views of the TEA model to select seeds that are reliable in one aspect yet informative for the other.

3.2.1 Localized Area

We first define the evolving entities as entities participating in newly added triples at snapshot k :

$$\mathcal{E}_{evo} = \mathcal{E}(\Delta\mathcal{F}^{(k)}), \quad \mathcal{E}'_{evo} = \mathcal{E}(\Delta\mathcal{F}'^{(k)}) \quad (1)$$

where $\mathcal{E}_{evo}$ (resp. $\mathcal{E}'_{evo}$) denotes the evolving entity set derived from the incremental temporal facts $\Delta\mathcal{F}^{(k)}$ of the source TKG (resp. $\Delta\mathcal{F}'^{(k)}$ of the target TKG).

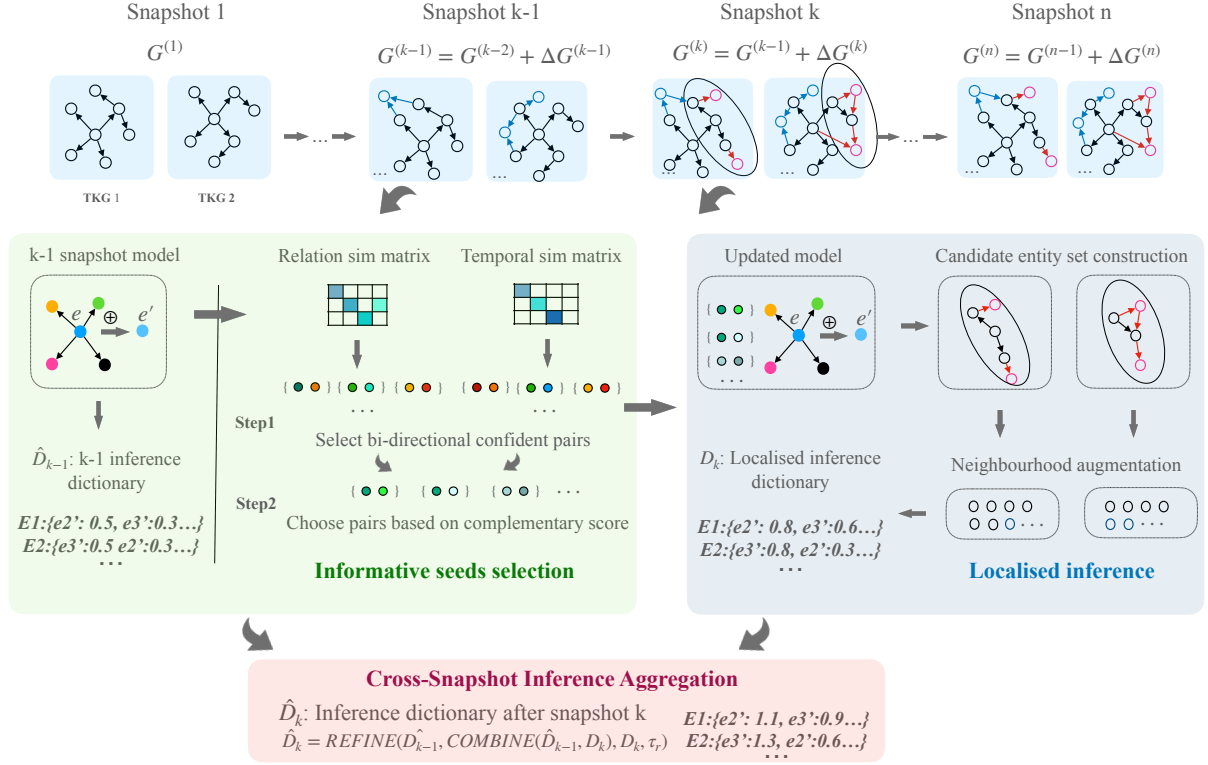


Figure 2: ELITE framework overview.

Our preliminary findings in Table 9 (Appendix D) reveal that newly added triples provide substantially stronger alignment signals within the evolving region than over the whole graph, owing to bounded information propagation in GNNs. We therefore restrict seed selection within the *localized area* — the subgraph induced by evolving entities and their one-hop neighborhood — as it concentrates propagation on the most recently changed structure. This localized strategy is further applied in our inference stage (Section 3.3), ensuring consistency between training and inference while avoiding redundant computation over unchanged regions of the TKG.

3.2.2 Dual-aspect Informative Seed Selection

TEA models (Liu et al., 2023; Li et al., 2025, 2026) leverage complementary relational and temporal information, analogous to multi-modal entity alignment. Motivated by active learning (Settles, 2009; Liu et al., 2021), we propose a *complementary uncertainty* criterion for dual-aspect seed selection, prioritizing entities with asymmetric confidence — certain in one modality while uncertain in the other. Such seeds carry richer signals than purely confident or uncertain pairs, enabling more effective fine-tuning across snapshots.

In particular, we follow Algorithm 1 to obtain

informative and complementary seeds from the localized area for model fine-tuning. Based on *bi-directional confident seeds* derived from both relational and temporal encoders (Li et al., 2025), we filter out *dual-aspect* confident pairs — those independently confirmed by both encoders, since they carry the least uncertainty for training the joint model. Among the remaining candidates, we then compute a *complementary modesty score* and select the Top- n pairs for each encoder.

The complementary modesty score of a candidate seed (e_i, e'_j) is calculated from the temporal score s_t and relational score s_r , as defined below:

$$\text{score}(s_t, s_r) = s_t \cdot g(s_r) + s_r \cdot g(s_t) \quad (2)$$

where the gating weights g is obtained from a Gaussian function centered at a target modesty value μ :

$$g(s) = \exp\left(-\frac{(s - \mu)^2}{2\sigma^2}\right) \quad (3)$$

Here, μ controls the target modesty level and σ governs the tolerance around it, where $s_t \cdot g(s_r)$ gates the contribution of s_t based on how close s_r is to μ , and vice versa. This formulation rewards two configurations: (i) one score is confidently high while the other remains modest near μ , allowing the confident score to contribute strongly; and (ii)

both scores are mutually modest around μ , enabling both terms to contribute fully. Conversely, it penalizes cases where both scores are simultaneously high, as the gates mutually suppress each other, or where both scores are low, yielding negligible contributions.

Given the newly selected informative seeds, we fine-tune the backbone TEA model for each snapshot using the original training objective.

3.3 Localized Inference and Aggregation

For TKG snapshot $\mathcal{G}^k$, we incrementally re-align a small portion of entities by integrating the accumulated historical evidence and the snapshot-specific predictions, to guarantee both alignment accuracy and inference efficiency.

3.3.1 Efficiency: Localized Entity Retrieval

The full-scale entity re-alignment over large TKG snapshots incurs substantial computational cost, as it requires computing similarity scores across all entity pairs at each snapshot. However, the evolving region generates the largest performance gain, owing to the enriched structural and temporal information introduced at each snapshot (Table 9 in Appendix D). This motivates us to constrain the inference to a localized evolving region, reducing computational overhead while concentrating the alignment capacity where it matters most.

Initial candidate entity set construction. Starting from the evolving region, we augment candidate entity sets of the source and target TKGs with their potentially matchable counterparts, to improve the recall of alignment predictions:

$$\begin{cases} \mathcal{E}_c = \text{Top}_1(\mathcal{E}'_{evo}) \cup \mathcal{E}_{evo} \\ \mathcal{E}'_c = \text{Top}_1(\mathcal{E}_{evo}) \cup \mathcal{E}'_{evo} \end{cases} \quad (4)$$

where $\mathcal{E}_c$ and $\mathcal{E}'_c$ denote the source and target candidate sets respectively. $\text{Top}_1(\cdot)$ denotes top-1 retrieved entity set. Extending the candidate set to incorporate Top-1 predicted counterparts is critical (as verified in Figure 6), since the potential matches may come from the non-evolving region of the counterpart TKG.

Bi-directional threshold-based filtering. Directly computing similarity over candidate entities is costly because of the uneven set size for source and target TKGs, and meanwhile introduces noise from weakly relevant pairs. We therefore retain only mutually consistent, high-confidence entity

pairs to balance recall and efficiency (as verified in Figure 7). We compute an initial similarity matrix:

$$\mathcal{S} = \text{Sinkhorn}(\mathbf{E}_{\mathcal{E}_c} \times \mathbf{E}_{\mathcal{E}'_c}^\top) \quad (5)$$

and retain only pairs satisfying mutual top-1 consistency and a minimum confidence threshold:

$$\mathcal{M} = \left\{ (e_i, e'_j) \left| \begin{array}{l} j = \arg \max_k \mathcal{S}(i, k), \\ i = \arg \max_k \mathcal{S}(k, j), \\ \mathcal{S}(i, j) > \tau_s \end{array} \right. \right\} \quad (6)$$

Where $\mathbf{E}_{(\cdot)}$ represents entity embeddings learned from both relational and temporal aspects, τ_s is the confidence threshold. Sinkhorn denotes Sinkhorn operator (Cuturi, 2013). The bi-directionality condition eliminates asymmetric matches, while τ_s discards low-confidence pairs. The refined source and target entity sets are then:

$$\begin{aligned} \mathcal{E}_f &= \{e_i \mid (e_i, e'_j) \in \mathcal{M}\}, \\ \mathcal{E}'_f &= \{e'_j \mid (e_i, e'_j) \in \mathcal{M}\} \end{aligned} \quad (7)$$

Neighborhood augmentation. The filtered sets $\mathcal{E}_f$ and $\mathcal{E}'_f$ are highly-precise but may under-cover entities whose alignment evidence is carried primarily by their neighborhood. We therefore augment each set with its one-hop neighbors:

$$\begin{aligned} \mathcal{E}_{fe} &= \mathcal{E}_f \cup \mathcal{N}_1(\mathcal{E}_f) \\ \mathcal{E}'_{fe} &= \mathcal{E}'_f \cup \mathcal{N}_1(\mathcal{E}'_f) \end{aligned} \quad (8)$$

where $\mathcal{N}_1(\cdot)$ denotes the one-hop neighborhood in the respective TKG. This enriches the subsequent similarity computation with broader relational and temporal context while keeping the inference region compact.

Localized inference dictionary generation. We apply the Sinkhorn operator to generate a locally optimized similarity matrix over the augmented candidate subgraph:

$$\hat{\mathcal{S}} = \text{Sinkhorn}(\mathbf{E}_{\mathcal{E}_{fe}} \times \mathbf{E}_{\mathcal{E}'_{fe}}^\top) \quad (9)$$

The top-ranked potential matches for each entity at snapshot k is then obtained as:

$$D_k(e_i) = \left\{ (e'_j, \hat{\mathcal{S}}(i, j)) \mid j \in \text{Top}_K(\hat{\mathcal{S}}(i, \cdot)) \right\} \quad (10)$$

By restricting Sinkhorn optimization to the localized candidate region, we substantially reduce computational cost while preserving alignment quality within the evolving area.

3.3.2 Accuracy: Cross-Snapshot Inference Aggregation

Having obtained localized inference results per snapshot, we aggregate alignment scores across snapshots into a coherent global inference. Two key issues arise: (i) how to effectively combine localized per-snapshot scores, and (ii) how to resolve conflicts when scores from previous snapshots disagree with the current snapshot on the top candidate for the same entity. We address both issues through a *dynamic threshold-based combination* strategy that maintains a running score estimate, accumulating evidence while selectively suppressing updates when accumulated confidence already exceeds that of the incoming snapshot signal.

Dynamic score combination. The running score estimate $\hat{D}_k$ is defined recursively: the first snapshot score is used directly as the initial estimate, and each subsequent estimate is formed by combining the previous running score with the current snapshot score via confidence-guided selection:

$$\hat{D}_k = \begin{cases} D_k & \text{if } k = 0 \\ \text{REFINE}(\hat{D}_{k-1}, \text{COMBINE}(\hat{D}_{k-1}, D_k), D_k, \tau_r) & \text{if } k > 0 \end{cases} \quad (11)$$

where COMBINE aggregates scores via summation, defined for entities present in both dictionaries, i.e., $e \in \hat{D}_{k-1} \cap D_k$:

$$\text{COMBINE}(\hat{D}_{k-1}, D_k)(e, e') = \hat{D}_{k-1}(e, e') + D_k(e, e') \quad (12)$$

and REFINE applies the confidence-guided update in the following.

Confidence margin and refinement. We define the *confidence margin* for entity e under score dictionary D as the score gap between its Top-1 and Top-2 candidates:

$$\delta(e, D) = D(e, e'_1) - D(e, e'_2) \quad (13)$$

where e'_1 and e'_2 are the Top-1 and Top-2 ranked candidates under D respectively. Since D_k is restricted to the current snapshot's localized candidate region (Eq. 4–9) while $\hat{D}_{k-1}$ accumulates entities across all prior snapshots, an entity e may fall outside D_k at a given snapshot; in this case its accumulated estimate is retained unchanged. Otherwise, the REFINE function retains the previous estimate when it carries higher confidence, and accepts the

new combination otherwise:

$$\hat{D}_k(e, e') = \begin{cases} \hat{D}_{k-1}(e, e') & \text{if } e \notin D_k \\ D_k(e, e') & \text{if } e \in D_k \text{ and } e \notin \hat{D}_{k-1} \\ \hat{D}_{k-1}(e, e') & \text{if } e \in D_k, e \in \hat{D}_{k-1}, \text{ and } \delta(e, \hat{D}_{k-1}) - \delta(e, D_k) > \tau_r \\ \hat{D}_k^c(e, e') & \text{otherwise} \end{cases} \quad (14)$$

where $\hat{D}_k^c = \text{COMBINE}(\hat{D}_{k-1}, D_k)$ denotes the combined score of entity e after merging the accumulated alignment dictionary $\hat{D}_{k-1}$ with the localized inference result D_k at snapshot k , and τ_r is a confidence threshold. If an entity falls outside the current snapshot's localized region, its accumulated estimate is carried forward unchanged. If an entity is newly introduced at snapshot k with no prior estimate, its score is taken directly from the current snapshot. Otherwise, if the accumulated confidence margin exceeds that of the incoming score by more than τ_r , we retain $\hat{D}_{k-1}$ without update; otherwise we accept $\hat{D}_k^c$. This prevents overwriting high-confidence accumulated predictions with less certain incoming scores, while ensuring entities with low running confidence continue to benefit from newly observed snapshot evidence.

4 Experiment

Datasets. We conduct experiments on two newly constructed evolving TEA benchmarks, detailed stats in Table 8. The first, YAGO-WIKI180K-E, is derived from the YAGO-WIKI180K dataset introduced by (Li et al., 2025) which is a large-scale TEA dataset derieve from YAGO and WIKI data with partial temporal coverage. The second, BETA-E, is based on the BETA dataset proposed by (Zeng et al., 2024), which is a fine-grained TEA benchmark incorporating hybrid temporal triples with date-level temporal granularity. We restructure both datasets to conform to chronological TKG evolving patterns, yielding YAGO-WIKI180K-E and BETA-E respectively.

Baselines. We adopt RCTEA (Li et al., 2026) as the backbone model, yet our ELITE framework can be applied to other dual-aspect TEA approaches (Liu et al., 2023; Li et al., 2025). We compare against three groups of baselines. The first is the *Full Retraining* group, where TEA models including STEA (Cai et al., 2022), TEA-GNN (Xu et al., 2021), DualMatch (Liu et al., 2023), HTEA (Li et al., 2025), and RCTEA (Li et al., 2026) are fully retrained from scratch at each snapshot, and Light-TEA (Cai et al., 2023), a training-free model, performs full inference at every snapshot. The second group applies the *Adjacency Matrices Update* strat-

Table 1: Overall TEA performance on YAGO-WIKI180K-E (2000 seeds), the p-value is the result of paired t-test between ELITE and their corresponding strong baselines on H@1.

Section	Model	Snapshots(Metric: H@1/5)									
		40	41	42	43	44	45	46	47	48	49
Full Retraining	TEA-GNN	.157/.224	.153/.220	.159/.226	.156/.225	.154/.223	.166/.234	.166/.236	.170/.238	.179/.247	.164/.233
	STEa	.177/.216	.180/.220	.183/.222	.185/.224	.188/.228	.190/.230	.192/.232	.194/.233	.196/.235	.193/.233
	LightTEA	.322/.406	.326/.413	.330/.415	.330/.415	.336/.421	.339/.423	.341/.425	.345/.429	.351/.423	.350/.473
	DualMatch	.323/.413	.323/.410	.332/.421	.334/.425	.335/.419	.338/.426	.339/.424	.346/.435	.345/.434	.342/.434
	HTEA	.317/.415	.315/.414	.322/.421	.325/.425	.329/.427	.326/.424	.330/.428	.331/.430	.333/.430	.328/.426
	RCTEA	.376/.471	.387/.480	.391/.484	.395/.487	.397/.488	.399/.491	.402/.492	.402/.492	.402/.491	.403/.493
Adjacency Matrices Update ¹	STEa	.177/.216	.180/.219	.183/.222	.184/.224	.189/.228	.189/.229	.192/.231	.193/.233	.195/.235	.194/.234
	HTEA	.317/.415	.320/.418	.323/.421	.323/.422	.328/.426	.327/.425	.328/.426	.330/.427	.331/.429	.330/.428
	RCTEA	.376/.471	.380/.474	.383/.477	.386/.479	.390/.482	.392/.484	.394/.486	.396/.487	.398/.489	.396/.488
Continual Learning	ContEA	.376/.471	.305/.385	.251/.306	.241/.295	.236/.288	.235/.287	.231/.280	.231/.279	.231/.278	.226/.273
	ELITE	.376/.471	.386/.478	.393/.484	.397/.487	.401/.490	.404/.492	.407/.494	.409/.496	.411/.497	.413/.498
	p-value	~	8.8e-3	6.9e-3	6.5e-3	6.5e-3	6.0e-3	5.9e-3	6.0e-3	6.1e-3	5.0e-3

Table 2: Overall TEA performance on BETA-E (10% seeds)

Section	Model	Snapshots(Metric: H@1/5)									
		10	11	12	13	14	15	16	17	18	19
Full Retraining	TEA-GNN	.241/.331	.263/.358	.287/.383	.311/.409	.333/.435	.367/.469	.389/.490	.410/.511	.433/.532	.448/.549
	STEa	.308/.395	.335/.426	.354/.448	.380/.474	.404/.500	.428/.525	.448/.545	.467/.567	.487/.582	.499/.596
	LightTEA	.358/.456	.381/.481	.376/.482	.411/.512	.458/.557	.448/.554	.476/.581	.528/.625	.501/.607	.553/.652
	DualMatch	.403/.488	.443/.528	.468/.552	.500/.582	.525/.610	.558/.639	.583/.664	.608/.688	.629/.707	.642/.719
	HTEA	.421/.523	.451/.554	.475/.579	.499/.598	.527/.621	.548/.641	.572/.665	.589/.681	.608/.692	.621/.709
	RCTEA	.446/.554	.477/.585	.496/.604	.521/.627	.547/.651	.572/.675	.593/.694	.615/.713	.631/.727	.643/.738
Adjacency Matrices Update	STEa	.308/.395	.333/.424	.351/.445	.374/.469	.395/.491	.416/.514	.436/.534	.452/.551	.470/.568	.480/.579
	HTEA	.421/.523	.446/.550	.462/.566	.484/.586	.501/.603	.519/.622	.536/.637	.550/.650	.563/.663	.571/.672
	RCTEA	.446/.554	.476/.582	.491/.598	.515/.620	.536/.640	.556/.661	.576/.678	.593/.695	.608/.709	.618/.718
Continual Learning	ContEA	.446/.554	.435/.530	.434/.524	.458/.546	.479/.564	.501/.583	.521/.633	.536/.644	.550/.657	.556/.664
	ELITE	.446/.554	.485/.588	.508/.608	.533/.629	.556/.651	.580/.671	.601/.689	.620/.705	.635/.718	.646/.727
	p-value	~	3.2e-3	1.9e-3	2.0e-3	2.2e-3	1.8e-3	1.9e-3	2.1e-3	2.0e-3	2.2e-3

egy (Yan et al., 2021), initializing from the first-snapshot model and updating adjacency matrix at each subsequent snapshot without model retraining. The third group comprises the *Continual Learning* baseline ContEA (Wang et al., 2022), validating our model’s performance against methods specifically designed for evolving entity alignment.

Evaluation Metrics. Following prior work, we adopt Hits@N (H@N) as our evaluation metric, which measures the proportion of correct entities ranked within the top-N retrieved candidates. Typically, we report H@1 and H@5 to evaluate alignment precision and recall performance respectively.

4.1 Main Results

RQ1: Can ELITE enhance evolving TEA performance over existing baselines? We compare ELITE against existing TEA models on YAGO-WIKI180K-E and BETA-E, using the 40th and 10th snapshots as the respective initial snapshots. As presented in Table 1 and 2, ELITE consistently

achieves state-of-the-art performance on $H@1$ and $H@5$, significantly outperforming both adjacency matrices update variants and ContEA across all snapshots. Notably, the performance gap widens as the number of snapshots increases, suggesting that ELITE benefits from accumulating snapshot evidence over time. On YAGO-WIKI180K-E, the local similarity update variant alone yields a 1.7% improvement over the best simple adjacency update baseline. On BETA-E, ELITE maintains a consistent advantage across all snapshots, achieving approximately 3% improvement over adjacency matrices update variants at the final snapshot while attaining competitive performance with state-of-the-art retraining baselines. This validates the effectiveness of ELITE in real-world evolving scenarios, with efficiency presented in the Section 4.3.

4.2 Efficacy of Informative Seed Selection

RQ2: Does the Dual-aspect Informative Seed Selection module improve alignment performance over confident seed selection variants? We investigate the effectiveness of our Dual-aspect Informative Seed Selection module by comparing it against three confident seed selection variants: **dual-conf**, which selects seeds based on bi-directional confidence using both relational and temporal informa-

¹TEA-GNN and DualMatch are excluded from the Adjacency Matrix Update setting due to their reliance on relation-specific attention mechanisms, which cause dimensional mismatches across evolving snapshots. LightTEA is excluded as its training-free inference paradigm is incompatible with the parameter update assumptions underlying the Adjacency Matrix Update setting.

tion jointly; **rel-conf**, which selects seeds based on relational confidence only; and **tem-conf**, which selects seeds based on temporal confidence only. All variants are evaluated against our informative seed selection using two encoder backbones.

As presented in Tables 3 and 4, our informative selection method consistently yields improvements across YAGO-WIKI180K-E and BETA-E datasets. Notably, the performance advantage over the second-best variant — **time-conf** on YAGO-WIKI180K-E — widens as the number of snapshots increases, suggesting that informative seed diversity becomes increasingly critical over longer evolving horizons. Interestingly, **dual-conf** performs the worst among all variants, indicating that incorporating dual-aspect confident seeds reduces the least uncertainty, which undermines the efficacy of the alignment model.

Table 3: Seeds selection on YAGO-WIKI180K-E (2000 seeds)

Model	Snapshots (H@1/H@5)				
	46	47	48	49	50
ELITE(dual-conf)	.396/.487	.397/.488	.398/.489	.399/.490	.400/.491
ELITE(rel-conf)	.396/.488	.397/.489	.398/.490	.399/.490	.400/.491
ELITE(time-conf)	.400/.490	.402/.491	.403/.492	.404/.493	.405/.494
ELITE	.407/.494	.409/.496	.411/.497	.413/.498	.415/.499

Table 4: Seeds selection on BETA-E (10% seeds)

Model	Snapshots (H@1/H@5)				
	15	16	17	18	19
ELITE(dual-conf)	.561/.663	.577/.678	.591/.692	.603/.704	.612/.713
ELITE(rel-conf)	.566/.670	.586/.688	.602/.703	.616/.715	.627/.725
ELITE(time-conf)	.569/.669	.587/.685	.603/.701	.617/.714	.627/.723
ELITE	.580/.671	.601/.689	.620/.705	.635/.718	.646/.727

4.3 Efficiency of Localized Inference

RQ3: Does the localized inference strategy achieve significant efficiency gains over full-graph inference? We evaluate the computational efficiency of ELITE against full-graph inference on YAGO-WIKI180K, a large-scale dataset comprising 180K entities designed to reflect real-world knowledge graph complexity. Rather than performing inference over all entities, ELITE selectively processes only the relevant evolving entities, motivated by the *local boosting* pattern whereby performance gains are largely concentrated in incremental regions of change.

We design two experiments to assess computational efficiency and entity retrieval accuracy at scale. As shown in Figure 3(a), the memory and

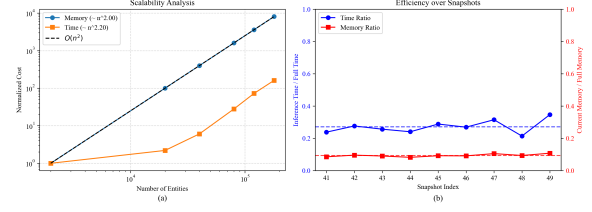


Figure 3: Scalability and efficiency on YAGO-WIKI180K-E. (a) Memory and runtime scaling as the number of entities increases, showing near-quadratic growth; (b) Relative inference time and memory usage across snapshots compared to full computation.

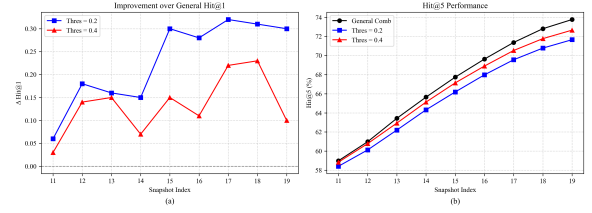


Figure 4: Hit@1/5 analysis across snapshots on BETA-E. (a) Hit@1 improvement over the general combination strategy under different threshold settings; (b) Hit@5 performance across snapshots, illustrating the trade-off between precision and recall under varying thresholds.

time consumption of full similarity matrix computation scales approximately as $\mathcal{O}(n^2)$ in practice. As presented in Figure 3(b), our localized inference requires on average only **27%** of the inference time and approximately **10%** of the memory of full matrix evaluation across snapshots, demonstrating substantial efficiency gains without sacrificing alignment accuracy.

4.4 Efficacy of Cross-Snapshot Aggregation

RQ4: Does the dynamic threshold-based combination strategy improve alignment performance over standard score combination? We analyze the cross-snapshot inference aggregation scores module on BETA-E dataset, and further examine the effect of the threshold-guided mechanism that suppresses low-confidence predictions based on the score gap between Top-1 and Top-2 candidates.

Specifically, we compare two variants and examine the effect of varying the threshold τ_r on $H@1$ and $H@5$, providing insights into the precision-recall trade-off involved. As shown in Figure 4(a), applying the threshold modification progressively improves $H@1$, as the model retains only high-confidence predictions from the score list. Notably, as $\tau_r \rightarrow 1$, the mechanism degrades to standard score combination without filtering. As shown

in Figure 4(b), an overly small threshold moderately reduces $H@5$ recall due to aggressive filtering limiting prediction diversity. However, as $H@1$ is the more critical metric under the one-to-one mapping assumption in practical entity alignment, these results further validate the effectiveness of our margin-based threshold mechanism in balancing precision and recall.

Table 5: Cold-start evaluation on BETA-E (Hits@1). ELITE initialised from snapshot 0 recovers warm-start performance by the final snapshot. Dashes denote snapshots preceding the warm-start initialisation point.

Snapshot	0	5	10	15	19
RCTEA (full retrain)	.161	.293	.446	.572	.643
ELITE (warm start=10)	–	–	.446	.580	.646
ELITE (cold start=0)	.161	.294	.437	.564	.638

Table 6: Cold-start evaluation on YAGO-WIKI180K-E (Hits@1). Cold-start ELITE tracks warm-start performance and exceeds full-retrain RCTEA from snapshot 10 onward. Dashes denote snapshots preceding the warm-start initialisation point.

Snapshot	0	10	20	30	40	49
RCTEA (full retrain)	.102	.183	.254	.321	.376	.403
ELITE (warm start=40)	–	–	–	–	.376	.413
ELITE (cold start=0)	.102	.211	.277	.334	.383	.412

4.5 Cold-Start Robustness

RQ5: Does ELITE’s effectiveness depend on a mature initialisation snapshot? ELITE is designed for the practical setting of an established KG that continues to grow, and our main experiments accordingly initialise from a mature snapshot (snapshot 10 of 20 on BETA-E, and snapshot 40 of 50 on YAGO-WIKI180K-E), avoiding snapshot 0, whose extremely sparse temporal facts provide a weak temporal-view signal. Because ELITE propagates pseudo-seeds across snapshots, an immature initial state yields noisier pseudo-seeds, and this noise could in principle compound over the alignment trajectory rather than wash out. To test this, we re-run ELITE from snapshot 0 on both datasets, keeping every other hyperparameter fixed, and compare against (i) RCTEA retrained from scratch at every snapshot and (ii) ELITE under its default warm start (Tables 5 and 6).

We find that cold-start ELITE converges to essentially the warm-start solution: by the final snapshot it recovers 98.8% of warm-start accuracy on BETA-E (.638 vs. .646) and is within a single point on YAGO-WIKI180K-E (.412 vs. .413), showing

that the early disadvantage from sparse initial supervision is transient rather than cumulative. It is also competitive with full retraining at a fraction of the cost, reaching 99.2% of RCTEA’s final accuracy on BETA-E (.638 vs. .643) and *exceeding* it on YAGO-WIKI180K-E (.412 vs. .403), where it leads at every snapshot from 10 onward. The consistency of these results across BETA-E (~50K entities) and YAGO-WIKI180K-E (~180K entities) suggests ELITE’s advantage does not depend on favourable initialisation and is not an artefact of scale.

5 Conclusion

In this paper, we proposed ELITE, an evolving TEA framework with three key designs: (i) dual-aspect complementary seed selection, which jointly exploits relational and temporal signals for per-snapshot fine-tuning; (ii) localized inference, which restricts re-alignment to selected evolving entities and reduces time and memory cost by around $4\times$ and $10\times$ respectively; and (iii) cross-snapshot inference aggregation, which accumulates alignment evidence while suppressing low-confidence updates. Extensive experiments on two real-world evolving TEA benchmarks, YAGO-WIKI180K-E and BETA-E, validate the superiority of ELITE.

6 Acknowledgements

This work was partially supported by the Australia Research Council under the Discovery Project (Grant No. DP230101753).

7 Limitation

Despite strong empirical performance, ELITE has several limitations. First, the bi-aspect seed selection module currently leverages only relational and temporal signals; extending it to multimodal settings incorporating visual or textual entity attributes represents a promising direction that we leave for future work. Second, the dynamic threshold τ_r requires dataset-specific tuning to optimally balance precision and recall, and an adaptive mechanism for automatic threshold estimation would further improve the robustness of the score combination strategy. Third, ELITE currently operates under a one-to-one entity mapping assumption, limiting its applicability to more general alignment settings involving one-to-many or many-to-many entity correspondences, which are common in real-world heterogeneous knowledge graphs.

References

- Luyi Bai, Longlong Zhou, Jingni Song, and Xuanxuan An. 2025. [An efficient and lightweight framework for unsupervised entity alignment between temporal knowledge graphs](#). *Information Fusion*, page 103821.
- Li Cai, Xin Mao, Youshao Xiao, Changxu Wu, and Man Lan. 2023. An effective and efficient time-aware entity alignment framework via two-aspect three-view label propagation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5021–5029.
- Li Cai, Xinnian Mao, Meirong Ma, Hao Yuan, Jianchao Zhu, and Man Lan. 2022. A simple temporal information matching mechanism for entity alignment between temporal knowledge graphs. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2075–2086.
- Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Jiayun Li, Wen Hua, Shiqi Fan, Fengmei Jin, Haiyang Jiang, and Xue Li. 2026. Rctea: Richness-guided co-training for temporal entity alignment. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 39295–39310.
- Jiayun Li, Wen Hua, Fengmei Jin, and Xue Li. 2025. Htea: Heterogeneity-aware embedding learning for temporal entity alignment. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 982–990.
- Yunfei Li, Lu Chen, Chengfei Liu, Rui Zhou, and Jianxin Li. 2023. Efficient and effective entity alignment for evolving temporal knowledge graphs. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 349–358. IEEE.
- Bing Liu, Harrison Scells, Guido Zuccon, Wen Hua, and Genghong Zhao. 2021. [ActiveEA: Active learning for neural entity alignment](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3364–3374.
- Songtao Liu, Rex Ying, Hanze Dong, Lanqing Li, Tingyang Xu, Yu Rong, Peilin Zhao, Junzhou Huang, and Dinghao Wu. 2022. Local augmentation for graph neural networks. In *International Conference on Machine Learning (ICML)*, pages 14054–14072.
- Xiaozhe Liu, Junyang Wu, Tianyi Li, Lu Chen, and Yunjun Gao. 2023. Unsupervised entity alignment for temporal knowledge graphs. In *Proceedings of the ACM Web Conference 2023*, pages 2528–2538.
- Burr Settles. 2009. Active learning literature survey.
- Megh Shukla. 2022. Bayesian uncertainty and expected gradient length-regression: Two sides of the same coin? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2367–2376.
- Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.
- Yuxin Wang, Yuanning Cui, Wenqiang Liu, Zequn Sun, Yiqiao Jiang, Kexin Han, and Wei Hu. 2022. Facing changes: continual entity alignment for growing knowledge graphs. In *International Semantic Web Conference*, pages 196–213. Springer.
- Chengjin Xu, Fenglong Su, and Jens Lehmann. 2021. Time-aware graph neural network for entity alignment between temporal knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8999–9010.
- Chengjin Xu, Fenglong Su, Bo Xiong, and Jens Lehmann. 2022. Time-aware entity alignment using temporal relational attention. In *Proceedings of the ACM Web Conference 2022*, pages 788–797.
- Yuchen Yan, Lihui Liu, Yikun Ban, Baoyu Jing, and Hanghang Tong. 2021. Dynamic knowledge graph alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4564–4572.
- Weixin Zeng, Jie Zhou, and Xiang Zhao. 2024. Benchmarking challenges for temporal knowledge graph alignment. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3103–3112.
- Runhao Zhao, Weixin Zeng, Wentao Zhang, Xiang Zhao, Jiuyang Tang, and Lei Chen. 2025. Towards temporal knowledge graph alignment in the wild. *arXiv preprint arXiv:2507.14475*.

A Gradient–Magnitude View of the Complementary–Modesty Score

We provide theoretical justification for the proposed informative seed selection module from a gradient-magnitude perspective.

A seed pair should be *reliable* for at least one view and, at the same time, inject a *large parameter update* into the view that remains uncertain. Below we show that the gating mechanism in Eq. (2) is a principled surrogate for the *expected gradient length* (EGL)(Shukla, 2022) that arises from this “reliability $\times$ update” trade-off.

Set-up. Let the temporal (T) and relational (R) encoders produce confidence scores $s_t, s_r \in (0, 1)$ for a candidate entity pair known to be a match

($y = 1$). Using per-instance cross-entropy losses $\ell(s_t) = -\log s_t$ and $\ell(s_r) = -\log s_r$, the squared gradient norm of a single sample with respect to the parameters θ_T of encoder T satisfies

$$\|\nabla_{\theta_T} \ell(s_t)\|^2 \propto \left(\frac{\partial \ell}{\partial s_t}\right)^2 = s_t^{-2}, \quad (15)$$

and an analogous expression holds for encoder R . This suggests that lower-confidence samples may induce larger parameter updates and therefore potentially provide stronger learning signals.

Reliability-aware weighting. However, in pseudo-label alignment, extremely low-confidence pairs are often unreliable and may introduce harmful supervision signals. In practice, the most informative seeds are usually those with *moderate confidence*: they remain sufficiently reliable while still capable of producing non-trivial parameter updates. To capture this property, we introduce a reliability-aware weighting function for the *other* view:

$$\begin{aligned} g(s_t) &= \exp\left[-\frac{(s_t - \mu)^2}{2\sigma^2}\right], \\ g(s_r) &= \exp\left[-\frac{(s_r - \mu)^2}{2\sigma^2}\right], \end{aligned} \quad (16)$$

where μ controls the preferred confidence region and σ controls its spread. The weighting function reaches its maximum when the partner view is moderately confident ($s \approx \mu$) and decreases when the score is either too low (unreliable) or too high (already certain and therefore less informative).

Reliability-weighted expected gradient utility. Using the reliability weighting from the complementary view, the expected gradient utility delivered to encoder T can be written as

$$\text{EGL}_T(s_t, s_r) = g(s_r) s_t^{-2}, \quad (17)$$

with the symmetric form

$$\text{EGL}_R(s_t, s_r) = g(s_t) s_r^{-2}. \quad (18)$$

This formulation reflects the intuition that a useful seed should not only induce meaningful parameter updates, but should also be supported by a sufficiently reliable complementary view.

A bounded surrogate. Directly ranking candidate pairs using Eqs. (17) and (18) is undesirable, since the inverse-square terms diverge as $s_t, s_r \rightarrow 0$, causing extremely noisy low-confidence pairs

to dominate the ranking. To improve robustness, we replace the unbounded inverse-square terms with bounded linear surrogates while retaining the reliability-aware modulation $g(\cdot)$.

Specifically, the two terms

$$s_t g(s_r) \quad \text{and} \quad s_r g(s_t)$$

respectively quantify the reliability-aware learning utility delivered to the temporal and relational encoders. The first term measures the expected contribution to the temporal view under the reliability support of the relational view, while the second term captures the symmetric contribution to the relational view. Summing the two terms therefore approximates the total mutual update utility across both views, encouraging candidate pairs that provide strong and reliable bidirectional supervision.

The resulting surrogate score is

$$s_t g(s_r) + s_r g(s_t), \quad (19)$$

which corresponds to Eq. (2). Therefore, Eq. (2) can be interpreted as a *bounded reliability-aware gradient proxy* that prioritises candidate pairs expected to induce strong and reliable mutual updates, while suppressing both noisy pairs (both low confidence) and redundant pairs (both excessively high confidence).

Behaviour of the gate.

- **One-high / one-modest:** if s_t is high and $s_r \approx \mu$, $g(s_r) \approx 1$ keeps the large s_t , driving a strong corrective gradient into encoder R .
- **Both modest:** $s_t \approx s_r \approx \mu$ implies $g(s_t) \approx g(s_r) \approx 1$, providing balanced updates to *both* views and sharpening their joint decision boundary.
- **Both high / both low:** the gated score (and thus the EGL surrogate) is near zero, so extremely easy or excessively noisy pairs are discarded.

Consequently, the complementary-modesty rule pursues seeds that maximise reliability $\times$ update, accelerating co-training and reducing confirmation bias in multi-view entity alignment.

B TEA and Evolving TEA Datasets

We evaluate our approach on two evolving temporal EA datasets constructed from YAGO-WIKI180K(Li et al., 2025) and BETA(Zeng et al.,

²BETA’s heterogeneous portion of temporal triples is measured at the date level.

Table 7: TEA Dataset Statistics: #Entities, # Relations, and #Facts indicate the number of entities, relations, and facts in two TKGs. Additionally, $|\mathcal{S}|$: number of ground-truth aligned entity pairs; $|\mathcal{P}|$: number of temporal points; λ_t : proportion of temporal triples; λ_{TH} : proportion of heterogeneous temporal triples.

Dataset	#Entities	#Relations	#Facts	$ \mathcal{S} $	$ \mathcal{P} $	λ_{t_1}	λ_{t_2}	λ_{TH}
YAGO-WIKI180K	187,987 187,977	32 261	924,935 1,636,020	187,977	1,000	.218	.265	.082
BETA	42,666 42,297	257 45	199,879 162,320	40,364	967	.645	.350	.390 ²

Table 8: Evolving TEA Dataset Statistics: #Entities, # Relations, and #Facts indicate the number of entities, relations, and facts in two TKGs. Additionally, n : number of snapshots $|\Delta\mathcal{E}^{(k)}|$: mean number of entities affected in evolving triples per snapshot; $|\Delta\mathcal{F}^{(k)}|$: mean number of triples added per snapshot; λ_{TE} : mean proportion of evolving triples per snapshot.

Dataset	#Entities	#Relations	#Facts	n	$ \Delta\mathcal{E}^{(k)} $	$ \Delta\mathcal{F}^{(k)} $	λ_{TE}
YW180K-E	187,987 187,977	32 261	924,935 1,636,020	50	5035 9210	4038 8683	.558 .722
BETA-E	42,666 42,297	257 45	199,879 162,320	20	4550 3697	5100 3389	5.36 3.67

2024), both of which feature heterogeneous temporal information — distinguishing them from homogeneous counterparts such as ICEWS, YAGO-WIKI50K, and YAGO-WIKI20K (Xu et al., 2021). Notably, YAGO-WIKI180K encompasses approximately 180K entities, making it a particularly challenging benchmark for model efficiency. Below, we describe the source dataset specifications and our newly constructed evolving variants.

YAGO-WIKI180K (Li et al., 2025). The YAGO-WIKI180K dataset is constructed using the same Wikidata source as the YAGO-WIKI50K and YAGO-WIKI20K datasets, with a key distinction in how temporal information is handled. The YAGO triples are extracted from YAGO3 (Suchanek et al., 2007). All available YAGO triples are first retrieved, followed by the extraction of temporal information from associated file records that contain timestamp details for each fact. This procedure enables the temporal information to be derived directly from the YAGO portion, rather than incorporating temporal data from the WIKI counterpart. Using available QID links, the YAGO and WIKI TKGs are aligned by selecting triples in which YAGO entities have corresponding counterparts in the WIKI graph. The resulting dataset includes approximately 19,000 shared entities, with 32 and 261 relations on the YAGO and WIKI sides, respectively, and includes a partial set of temporal triples. Compared to other existing TEA datasets, this construction process yields nearly four times as many entities. Furthermore, due to the independent sourcing of temporal information from the two KGs, the dataset exhibits temporal heterogeneity in 8.2% of the aligned triples. This level of het-

erogeneity more accurately reflects the variation commonly encountered in real-world settings.

BETA (Zeng et al., 2024). Unlike all the other datasets which standardizes timestamps to year-level granularity, BETA retains multi-granular temporal information, with 28% of temporal facts in the Wikidata subset and 14% in the YAGO subset containing precise month-day annotations. This fine-grained temporal data introduces more diverse information in alignment tasks. In terms of entity distribution, BETA exhibits a more realistic spread: while most entities in YAGO-WIKI50K are associated with a single dominant temporal relation (e.g., `member_of_sports_team`), BETA contains entities with no temporal facts as well as those linked to multiple temporal relations similar to YAGO-WIKI180K. This distribution supports the design of new alignment scenarios beyond the original single-relation setting. Additionally, relation frequency in BETA is more balanced: whereas `member_of_sports_team` and `plays_for` account for over 98% of all quadruples in YAGO-WIKI50K, the top five relations in BETA collectively contribute to less than 35%, reflecting a broader and more domain-general knowledge structure. These characteristics make BETA a more challenging and representative benchmark for the TEA task.

YAGO-WIKI180K-E. This evolving dataset is derived from YAGO-WIKI180K (Li et al., 2025), a large-scale temporal knowledge graph featuring partial temporal annotations. To construct the evolving version, we first anchor the static triples associated with each snapshot in the subject KG, then identify the corresponding temporal end date for the object KG triples to be incrementally added.

This procedure ensures that each snapshot receives an approximately uniform number of new triples, simulating real-world knowledge graph evolution. The resulting dataset comprises 50 snapshots, with a mean of 4,038 and 8,683 incremental triples per snapshot for the subject and object KGs respectively, involving approximately 5,035 and 9,210 entities. The incremental triple ratios are 0.558 and 0.722 respectively.

BETA-E. The BETA-E dataset is constructed following a procedure analogous to YAGO-WIKI180K-E. Specifically, we anchor a fixed number of incremental triples for the subject KG per snapshot, then mirror the corresponding time interval to select the counterpart incremental triples for the object KG, forming a temporally aligned evolving dataset based on BETA(Zeng et al., 2024). The resulting dataset comprises 20 snapshots, with mean incremental triples of 5,100 and 3,389 per snapshot for the subject and object KGs respectively, involving approximately 4,550 and 3,697 entities. The incremental triple ratios are 5.36 and 3.67 respectively.

C Parameter Setting

We detail the parameter settings used in our experiments. Our implementation is based on the original RCTEA backbone.

For both YAGO-WIKI180K-E and BETA-E, we set the embedding dimension to $d = 50$, the number of attention layers to $L = 1$, the batch size to $b = 512$, and the dropout rate to $drop = 0.3$. The model is optimized using RMSprop with a learning rate of $\eta = 0.005$. The number of training iterations for the first snapshot is set to $E_{first} = 15$, and the fine-tuning iterations for subsequent snapshots are set to $E_{fine-tuning} = 6$.

For informative seed selection, we set $\sigma = 0.15$ and $\mu = 0.8$ for both datasets, with bi-directional threshold $\tau_s = 0.8$. For threshold fusion, we set $\tau_r = 0.4$ and $\tau_r = 0.3$ for BETA-E and YAGO-WIKI180K-E, respectively. The number of fine-tuning seeds is fixed at $n = 1000$ per encoder, and the top- k candidate set size for score dictionary construction is set to $k = 100$.

The Top-1 retrieval operation in Eq. 4 is conducted using the Faiss (Johnson et al., 2019) framework to enable efficient large-scale nearest-neighbor embedding retrieval.

All experiments are conducted on a Linux cluster equipped with an AMD EPYC 3 Milan CPU, an

NVIDIA H100 GPU, and 1,500 GB of RAM.

The experiments are run over 5 times and the mean is reported. All remaining experiments are performed using the original implementations and their respective hyperparameter settings.

D Preliminary Study

Table 9: Preliminary study on YAGO-WIKI180K-E (last 5 snapshot transitions)

Metric	45→46	46→47	47→48	48→49
Overall Alignment (H@1/5)	.457/.460	.460/.463	.463/.466	.466/.465
Evolving Region (H@1/5)	.898/.908	.865/.874	.853/.862	.869/.868
Non-evolving Region (H@1/5)	.269/.270	.269/.270	.266/.266	.273/.272
Consistent Alignment Cases	21477	21646	22368	21853
Improved Evolving Entity Cases	140	125	139	−13
Improved Non-evolving Entity Cases	20	7	4	−15
Total Improved Cases	160	132	143	−28

Preliminary Study: Evolving In-Region vs. Out-of-Region Entity Alignment. We conduct a preliminary study on YAGO-WIKI180K-E using RCTEA adjacency matrices update variant to analyze where performance gains are concentrated across snapshots. Specifically, we conduct alignment on a random subset and decompose alignment performance into two disjoint subsets: the *evolving(localized) region*, comprising evolving entities and their immediate neighborhood, and the *non-evolving region*, the remaining entity set, i.e., all entities excluding those contained in the evolving region.

As shown in Table 9, a substantial performance gap exists between the evolving and non-evolving regions. Here, the evolving region consists of entities involved in newly added triples at the current snapshot, while the non-evolving region is defined as the remaining entities outside the evolving region, i.e., all entities excluding those associated with evolving triples. The evolving region consistently achieves significantly higher alignment accuracy (H@1: 85%–90%) compared to the non-evolving region (H@1: 26%–27%), indicating that newly introduced triples provide substantially stronger and more informative alignment signals. Furthermore, improvement cases are overwhelmingly concentrated in the evolving region. Across snapshots 45→46 through 47→48, the evolving region contributes 140, 125, and 139 improved cases, whereas the non-evolving region contributes only 20, 7, and 4 cases, respectively.

Meanwhile, the number of consistent alignment cases is substantially larger than the number of improved cases, suggesting that only a relatively

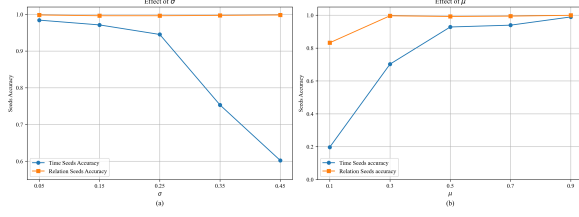


Figure 5: Sensitivity analysis of the Gaussian gating parameters σ and μ on seed accuracy. (a) Influence of the variance parameter σ on temporal and relation seed accuracy. As σ increases, temporal seed accuracy decreases significantly, while relation seed accuracy remains consistently high. (b) Influence of the target modesty value μ on seed accuracy. Increasing μ substantially improves temporal seed accuracy and maintains stable relation seed performance, with the best overall seeds accuracy achieved at higher μ values.

small portion of entities are affected by incremental graph evolution compared to the overall graph scale.

These observations motivate our core design: instead of repeatedly performing full-graph inference at every snapshot, we focus both fine-tuning and inference on the evolving region, where alignment gains are most significant. This directly supports the effectiveness of our proposed localized seed selection and localized inference framework.

E Effect of Parameters in Informative Seed Selection

In this section, we examine the effect of σ and μ on seed selection accuracy at snapshot 41 of the YAGO-WIKI180k-E dataset, as defined in Eq. 2. As shown in the figure, the bi-directional relational seeds maintain consistently high accuracy even as μ decreases, suggesting that the bi-directional consistency filter alone is sufficient to ensure high-confidence relational seed selection. In contrast, the temporal seeds exhibit a clear degradation trend: as μ decreases and σ increases, seed accuracy drops substantially. This observation further validates that relational signals are inherently more robust than temporal signals, and that temporal seed quality is more sensitive to the choice of scoring hyper-parameters.

F Benefit of Including Top-1 Counterpart in Candidate Set Construction

We further analyze the impact of the top-1 retrieval strategy in Eq. 4. As illustrated in the Figure. 6, incorporating top-1 retrieval entities consistently

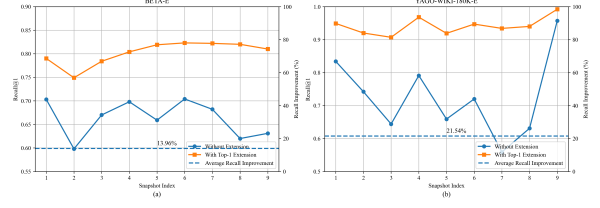


Figure 6: Impact of the Top-1 extension strategy on localized inference recall across evolving snapshots. (a) Recall comparison on the BETA-E dataset between localized inference without extension and the proposed top-1 extension mechanism. The extension strategy consistently improves alignment recall, achieving an average improvement of 13.96%. (b) Recall comparison on the YAGO-WIKI-180K-E dataset, where the proposed extension mechanism further enhances localized inference performance with an average recall improvement of 21.54%.

improves recall across evolving snapshots compared to only use $\mathcal{E}'_c$ as the target entities. Specifically, the proposed extension mechanism achieves an average recall improvement of approximately 14% on BETA-E and 21% on YAGO-WIKI-180K-E. These results validate our enrichment design, demonstrating that complementing evolving entities with $\mathcal{E}'_c$ substantially enhances alignment coverage and overall recall performance in the target graph.

G Analysis of Bi-directional Threshold Localized Entity Retrieval

We further examine the efficacy and rationale of our bi-directional threshold localized entity retrieval strategy. Specifically, we analyze the confidence threshold introduced in Eq. 6, which governs the trade-off between recall and memory consumption. As demonstrated empirically, our threshold-based method incurs only a marginal recall reduction of $\sim 2\%$ compared to full similarity matrix computation, while reducing the memory required for computing the final similarity matrix in Eq. 9 by approximately 60%. These results validate the efficiency and practicality of our localized retrieval strategy in large-scale TKG alignment settings.

H Backbone Generality

ELITE's incremental update mechanism is designed to be decoupled from the underlying temporal entity alignment (TEA) architecture, rather than tied to the specific backbone used in our main experiments. To test this, we integrate ELITE with

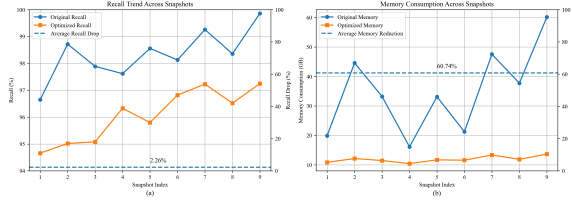


Figure 7: Trade-off analysis between recall performance and memory consumption across temporal snapshots. (a) Comparison of the original and optimized recall results over nine snapshots. The optimized approach introduces only a minor average recall decrease of 2.26%, indicating that alignment performance is largely preserved. (b) Comparison of memory consumption between the original and optimized methods. The optimized approach substantially reduces memory usage, achieving an average memory reduction of 60.74% across all snapshots.

HTEA (Li et al., 2025), an independently developed TEA model, and evaluate on BETA-E under the same snapshot-10 initialisation as our main setting. We compare HTEA+ELITE against (i) HTEA retrained from scratch at every snapshot, (ii) HTEA combined with a naive Adjacency Matrix Update baseline, and (iii) an HTEA-based ContEA (Table 10).

HTEA+ELITE consistently outperforms both the Adjacency Matrix Update baseline and ContEA at every snapshot, with its margin over the naive baseline widening from +0.021 to +0.040 by the final snapshot, and it recovers 98.4% of full-retraining accuracy (.611 vs. .621) without ever being retrained from scratch after initialisation. This confirms that ELITE’s incremental mechanism transfers to a different TEA backbone with no architecture-specific modification, and that its benefit is not an artefact of the backbone it was originally developed on.

Table 10: Backbone generality on BETA-E (Hits@1). HTEA+ELITE consistently outperforms the Adjacency Matrix Update baseline and ContEA, and recovers 98.4% of full-retraining accuracy at the final snapshot.

Snapshot	10	13	16	19
HTEA (full retrain)	.421	.499	.572	.621
HTEA (Adjacency Matrix Update)	.421	.484	.536	.571
ContEA (HTEA-based)	.421	.451	.512	.550
HTEA+ELITE	.421	.505	.557	.611

I Pseudo-Label Quality

ELITE’s incremental updates depend on the quality of the pseudo-seeds propagated across snap-

shots, and a natural seed-selection strategy is to prioritise *confident* predictions — those the current model scores most highly, whether by relational or temporal signal. We test whether confidence is in fact the right criterion by comparing two confidence-based baselines, Confident (rel) and Confident (time), against our informative seed-selection strategy. We report average seed precision and downstream model accuracy on BETA-E over the final 10 snapshots (Table 11).

Both confidence-based strategies reach near-saturation seed precision (.993 and .984 respectively), but this precision is misleading: because these seeds are predictions the model is already certain of, they carry near-zero gradient signal and contribute little to training (Appendix A). Our informative strategy trades some of this precision away — starting at .788 and rising to .901 as the model improves, for an average of .858 — in exchange for seeds that are actually useful for learning. This trade-off pays off: informative seeds yield the highest model accuracy of the three strategies (.561, ranging .446 → .646), exceeding both Confident (rel) (.549, .446 → .627) and Confident (time) (.550, .446 → .627). This confirms that seed precision alone is an incomplete proxy for seed utility, and that selecting for informativeness rather than confidence is what drives ELITE’s downstream gains.

Table 11: Pseudo-label quality on BETA-E, averaged over the last 10 snapshots. Confidence-based seed selection reaches near-saturation precision but yields near-zero gradient signal; our informative seeds trade precision for training utility and achieve the highest model accuracy.

	Confident (rel)	Confident (time)	Informative (ours)
Seed precision	.993	.984	.858 (.788 → .901)
Model accuracy	.549 (.446 → .627)	.550 (.446 → .627)	.561 (.446 → .646)

J Score Fusion Comparability

ELITE maintains a global score matrix over all entity pairs, and the localised inference performed at each snapshot updates only the scores for candidate pairs affected by that snapshot’s evolution. Because D_k (the current scores for candidate pairs at snapshot k) and $\hat{D}_{k-1}$ (the historical scores accumulated up to snapshot $k - 1$) are computed on the same underlying score scale, the two are directly comparable during cross-snapshot fusion — a property that a purely rank-based alternative would not preserve, since ranks alone discard the magnitude

of model confidence.

To test whether this comparability matters in practice, we compare our threshold-based RE-FINE fusion against Reciprocal Rank Fusion (RRF, $v=60$) on BETA-E over the final 10 snapshots (Table 12). Score fusion substantially and consistently outperforms rank fusion at every snapshot, with the gap widening from +0.053 at snapshot 11 to +0.209 by snapshot 19. This confirms that preserving fine-grained confidence information, rather than reducing scores to ranks, is central to ELITE’s cross-snapshot fusion design.

Table 12: Score fusion vs. rank-based fusion on BETA-E (Hits@1), last 10 snapshots. Score-based fusion (RE-FINE) substantially outperforms Reciprocal Rank Fusion, with the margin widening over time.

Snapshot	11	13	15	17	19
Rank fusion (RRF)	.432	.396	.400	.418	.437
Score fusion (ours)	.485	.533	.580	.620	.646

K Generalisation to a Standard TEA Benchmark

Our two original benchmarks are constructed to be evolving and heterogeneous, which makes them challenging but leaves open whether ELITE’s gains transfer to a more conventional setting. To test this, we build a third evolving benchmark, YAGO-WIKI50K-E, from YAGO-WIKI50K — a widely-used existing TEA dataset with homogeneous, full temporal coverage, and consequently an easier setting than our two original benchmarks — using the same chronological construction procedure described in Appendix B. We compare ELITE against ContEA, RCTEA retrained from scratch at every snapshot, and RCTEA with a naive Adjacency Matrix Update (Table 13).

ELITE outperforms the full-retrain variant at every snapshot, with the largest gain early (+0.092 at snapshot 5) narrowing to a smaller but consistent lead by the end (+0.015 at snapshot 19). We hypothesise the early gain reflects informative pseudo-seed selection giving the incrementally fine-tuned model a head start over retraining from scratch at each snapshot. These results confirm that ELITE’s effectiveness generalises beyond our two original benchmarks to a standard, widely-used TEA dataset.

Table 13: Evaluation on YAGO-WIKI50K-E (Hits@1). ELITE outperforms full retraining at every snapshot, with the largest gain early and a smaller but consistent lead by the final snapshot.

Snapshot	0	5	10	15	19
ContEA	.669	.626	.611	.755	.831
RCTEA (full retrain)	.669	.667	.718	.848	.940
RCTEA (Adjacency Matrix Update)	.669	.406	.444	.513	.515
ELITE	.669	.759	.802	.893	.955

L Fine-Grained Evaluation

To understand where ELITE’s gains on the evolving region originate, we further split it into *newly-occurred entities* and *old entities with new facts*, and report both against the non-evolving region, averaged over the final five snapshots (45–49) of YAGO-WIKI180K-E (Table 14). New entities account for only 0.3% of the evolving region (148 of 56,850 mean count), so the overall evolving-region score is driven almost entirely by old entities with new facts, whose Hits@1 (.877) is nearly identical to the evolving region as a whole (.876). The non-evolving region trails substantially at .197, a gap suggesting that newly added facts provide richer structural and temporal context that benefits alignment even for entities that were already present.

Table 14: Fine-grained breakdown on YAGO-WIKI180K-E (Hits@1), averaged over snapshots 45–49. Old entities with new facts drive the evolving-region score; the non-evolving region trails substantially.

	New entities	Old entities w/ new facts	Evolving region (overall)	Non-evolving region
Mean count	148	56,702	56,850	128,601
Mean H@1	.494	.877	.876	.197

M End-to-End Efficiency

We decompose ELITE’s runtime on YAGO-WIKI180K-E over snapshots 40–49 into its four stages (Table 15). Localized inference — similarity computation followed by Sinkhorn normalization — dominates end-to-end cost at 82.9% of total runtime, and for the same reason also dominates peak memory, since pairwise similarity computation scales as $O(n^2)$ in candidate set size; fine-tuning and cross-snapshot fusion, by contrast, operate on much smaller tensors and together account for under 5% of runtime. This confirms the necessity of localized inference as a design choice: system-level efficiency gains come primarily from reducing the candidate set size for inference, rather than from the fine-tuning or fusion stages.

Table 15: Runtime decomposition of ELITE on YAGO-WIKI180K-E, averaged over snapshots 40–49. Localized inference dominates both runtime and peak memory cost.

Stage	Mean time (units)	Percentage
Seed selection	217	12.9%
Fine-tuning	66	3.9%
Localized inference (Similarity + Sinkhorn)	1400	82.9%
Cross-snapshot fusion	5	0.3%

N Sinkhorn Operator

The Sinkhorn operator utilizes a fast and effective way to compute the assignment problem. It iteratively normalizes rows and columns to generate results in a doubly stochastic matrix, as follows:

$$\begin{aligned}
 \text{Sinkhorn}^0(S) &= \exp(S), \\
 \text{Sinkhorn}^m(S) &= \mathcal{N}_c(\mathcal{N}_r(\text{Sinkhorn}^{m-1}(S))), \\
 \text{Sinkhorn}(S) &= \lim_{x \rightarrow \infty} \text{Sinkhorn}^m(S)
 \end{aligned}$$

where $\mathcal{N}_r(S) = S \oslash (S \mathbf{1}_N \mathbf{1}_N^T)$, $\mathcal{N}_c(S) = S \oslash (\mathbf{1}_N \mathbf{1}_N^T S)$ are row and column-wise normalization operators of a matrix, $\oslash$ represents the element-wise division, and $\mathbf{1}_N$ is a column vector of ones.

O Related Work

Temporal Entity Alignment. Early TEA models such as TEA-GNN (Xu et al., 2021) and TREa (Xu et al., 2022) incorporate temporal point embeddings for temporal representation, yet treat temporal information analogously to relational data, overlooking its unique structural properties. Subsequent methods explore temporal information as an explicit attribute: STEA (Cai et al., 2022) encodes it via a temporal dictionary, DualMatch (Liu et al., 2023) integrates a temporal encoder into message propagation, and MGTEA (Zeng et al., 2024) exploits fine-grained temporal cues through a similar strategy. However, these approaches incur substantial computational overhead due to the construction and processing of temporal similarity matrices. LightTEA (Cai et al., 2023) introduces a lightweight alignment framework, though its temporal module yields only marginal improvement (Xu et al., 2021). In a similar pursuit of efficiency, ELTEA (Bai et al., 2025) combines relationship-aware graph attention with a Jaccard-based temporal similarity matrix, casting alignment as an optimal transport problem and generating confident pseudo-pairs to enable unsupervised training. HTEA (Li et al., 2025) further proposes an L_1 -weighted feature selection matrix

and a temporal attention module to enhance temporal propagation, alongside a heterogeneity-aware module to mitigate noise from heterogeneous temporal triples. Extending this heterogeneity concern to real-world, cross-source settings, Zhao et al. (2025) identify multi-scale temporal entanglement and structural imbalance as obstacles under-addressed by prior TEA methods, and propose a hypergraph retrieval-augmented framework together with new benchmarks to evaluate alignment in the wild. RCTEA (Li et al., 2026) proposes a richness attention mechanism to selectively propagate informative signals, paired with a joint neighbourhood consensus module to enforce alignment consistency across local neighbourhoods.

Evolving Entity Alignment. Several pioneer works have explored evolving EA under simulated snapshot settings (Wang et al., 2022; Yan et al., 2021). Specifically, Yan et al. (2021) proposes updating only the current adjacency matrix while maintaining a training-free paradigm for next-snapshot inference. ContEA (Wang et al., 2022) simulates evolving patterns via a confidence-guided seed selection module for snapshot fine-tuning, combined with a gating-based parameter update mechanism that selectively updates evolution-related embeddings. ITEA (Li et al. (2023)) instead adopt a teacher-student paradigm for incremental alignment, distilling a graph-attention teacher into a lightweight, sampling-based graph convolutional student to keep pace with the continual stream of new entities and relations characteristic of evolving TKGs; however, their evaluation reports only two incremental periods, offering a coarser view of evolving performance than the fine-grained, snapshot-by-snapshot analysis we conduct.