Entity Alignment for Multimodal Temporal Knowledge Graph

Abstract. Entity alignment (EA) aims to identify equivalent entities across distinct knowledge graphs (KGs). While existing methods leverage either temporal, structural, or visual information independently, they overlook the synergistic integration of multimodal data. This limitation leads to the absence of dedicated multimodal temporal knowledge graph (MTKG) benchmarks. Current MTKGs exhibit imperfections and are continuously updated; thus, we aim to address these limitations through alignment techniques. Moreover, most state-of-the-art approaches rely on simplistic fusion strategies—such as direct concatenation, averaging, or element-wise addition—to combine multimodal features, resulting in the incomplete utilization of information, leading to the alignment effect falling short of expectations. To address these gaps, we introduce ICEWS+, a novel multimodal temporal KG entity alignment dataset extending the ICEWS benchmark, and propose DynEA: a contrastive learning-based fusion framework enhanced by adaptive attention mechanisms. Empirical evaluations on ICEWS+ demonstrate that DynEA achieves a Hits@1 score of 95.41%, significantly outperforming existing methods.

Keywords: Knowledge Graphs · Entity Alignment · Multimodal Temporal Knowledge Graph Dataset · Contrastive Learning.

1 Introduction

The rapid advancement of artificial intelligence has accelerated research progress in domains such as information retrieval, intelligent question answering, and recommendation systems, accompanied by exponential growth in cross-domain resources and data volumes. To enhance resource management and utilization efficiency, KGs have emerged as a critical solution. Formally introduced in Google's 2012 official blog post[1] and successfully deployed in its search platform, KGs have since gained significant scholarly attention. By systematically defining and interlinking real-world concepts, entities, and their relationships, KGs structure web information using cognitive models aligned with human reasoning patterns. This framework endows machines with semantic comprehension capabilities for massive datasets, enabling deep information association, intelligent reasoning, and knowledge discovery—thereby establishing a cognitive infrastructure for intelligent applications.

[1] https://blog.google/products/search/introducing-knowledge-graph-things-not/

Current knowledge graph research, while having achieved independent fusion of multimodal data (e.g., images) and temporal information[2], still exhibits significant gaps in multimodal temporal collaborative modeling. MTKG can effectively characterize the dual properties of visual representations of entity attributes and their temporal evolution, significantly enhancing knowledge graphs' modeling capabilities for dynamic complex scenarios in the real world.

However, most existing MTKGs originate from disparate sources or monolingual contributions, constraining knowledge coverage. Consequently, matching and synchronizing independently constructed KGs to provide complementary information for *Natural Language Processing* (NLP) tasks is imperative[2]. *Knowledge fusion* addresses this challenge by integrating heterogeneous information sources. Early fusion research focused on conceptual-level mappings between KGs, while contemporary efforts prioritize data-level EA due to increasing data volumes[2].

The theoretical significance of EA manifests in two dimensions: Overcoming single-KG coverage limitations through multi-source integration, expanding downstream application scope[10]. (e.g., question answering systems) Reducing large-scale domain-specific KG construction costs (e.g., healthcare) by fusing precise smaller KGs, balancing scale and accuracy[23].

Most state-of-the-art EA approaches[16, 17] rely on simplistic fusion strategies—such as direct concatenation, averaging, or element-wise addition—to combine multimodal features. Critical analysis of existing EA research reveals three fundamental limitations:

- Dataset deficiency, with no existing EA datasets simultaneously incorporating multimodal and temporal information;
- Feature synergy neglect, where current techniques prioritize character, attribute, and relation features while underutilizing complementary information within and across multimodal KG modalities [38];
- Fusion simplicity, where prevalent fusion methods (concatenation, summation, averaging) lack adaptive weight selection mechanisms.

To address these gaps, this paper constructs a novel MTKG entity alignment dataset integrating multimodal attributes with temporal information, and develops *DynEA*—an innovative alignment framework leveraging contrastive learning, attention mechanisms, and graph neural networks. The *DynEA* framework integrates three synergistic modules: a *multimodal temporal data preprocessing module* that transforms heterogeneous data modalities—including entity names, relations, temporal attributes, and unstructured visual content—into unified embedding representations; a *contrastive learning-based fusion weight training module* that dynamically optimizes cross-modal integration weights through attention mechanisms and contrastive learning; and a *graph neural network-based encoder-decoder module* where specialized temporal and relational encoders process embeddings, reformulating entity alignment as a weighted graph matching problem to fuse spatiotemporal features, ultimately generating the alignment probability matrix through sigmoid-activated decoding. Our principal contributions include:

- Pioneering Dataset: Constructs the multimodal-temporal KG entity alignment benchmark with generalized construction methodology applicable across domains;
- Synergistic Feature Integration: Augments traditional EA methods by jointly incorporating visual features and timestamp-based temporal information to significantly enhance alignment accuracy;
- Advanced Fusion Algorithm: Proposes *DynEA* (Dynamic Embedding Alignment) featuring dynamic modality fusion with temporal modeling, validated through rigorous experiments for superior performance.

2 Related Work

Conventional EA methods. Existing EA methods can be divided into three types. Translation-based methods, like MTransE [6], BootEA [28], and AlignE [27], founded on TransE's framework [1], excel in knowledge representations. Graph Neural Networks(GNNs), exemplified by GCN [14], mark a notable advance in EA by aggregating neighborhood information to generate entity embeddings. GCN-Align [34], RDGCN [8], and Dual-AMN [21] exemplify GNN-based EA methods, utilizing GCN for modeling structure information and learning entity embeddings. Recent GNN-based methodologies, e.g., TEA-GNN [36], TREA [35], and STEA [3], have integrated temporal data, underscoring its significance in EA. Other approaches, such as Fualign [32], Simple-HHEA [12], and BERT-INT [31] address the heterogeneity in KGs by utilizing side information.

Temporal knowledge graph EA methods. Xu et al.[36] first studied temporal knowledge graph EA, treating timestamps as link attributes instead of discretizing temporal graphs into snapshots. They used time-aware attention to fuse information and optimize training. Tem-EA combines LSTM with GCN structural embeddings for alignment[25], while Sun et al. enhance graph attention with temporal modeling to learn entity embeddings[26]. Cai et al.[3] argue time labels need no separate representation, using a simple GNN with temporal matching for unsupervised alignment. Liu et al. generate initial labels via shared time information, fusing temporal and relational data with label-free encoders[17]. Another study mines entity evolution through time contexts but uses attribute-based temporal info[18].

Multimodal knowledge graph EA methods. Since Liu et al.[19] introduced visual modality into multimodal knowledge graph (MMKG) EA, this research direction has gained increasing attention alongside advances in multimodal learning. Chen et al.[4] fuse modality representations to minimize entity embedding distances; Liu et al.[16] apply attention mechanisms for modality weighting; Chen et al.[5] integrate visual features to guide relation learning and select key attributes; while Lin et al.[5] enhance intra-modal learning through contrastive learning and KL divergence. However, existing methods critically overlook dynamic inter-modal effects between entities and practical challenges, including KG noise, intra-modal feature discrepancies (e.g., node degree variations), and inter-modal preferences (e.g., modality absence or imbalance).

Our method fully integrates image information, temporal information, entity information, and structural information. It also utilizes technologies such as graph neural networks in deep learning to assist in alignment. Through intra-modal and inter-modal contrastive learning as well as attention mechanisms, it overcomes the defects of existing EA techniques and establishes a significant benchmark method for future MTKGs.

3 Construction Methodology for Multimodal Temporal Knowledge Graph Entity Alignment Datasets

Traditional knowledge graphs fail to capture dynamics and integrate multi-dimensional information. Advances in network technology enable access to multi-modal data and temporal recording via timestamps, motivating the development of *multimodal temporal knowledge graph EA*. This technique integrates temporal dynamics and multimodal complementarity to overcome static constraints and unimodal information gaps, enriching semantic representations for EA and knowledge completion.

To address the lack of public datasets, we construct **ICEWS+**. This benchmark supports algorithm validation and advances applications such as dynamic knowledge reasoning and question answering systems, meeting high-quality knowledge service demands in domains like healthcare and military operations.

3.1 Dataset Construction Process

After acquiring raw structured event data from the official ICEWS database[11], initiate the cleaning process. Through a machine cleaning layer, resolve the root directories of entity names and iteratively query entity names to generate a cleaned entity name list.

Based on the cleaned entity list, deploy a Playwright automation cluster for visual data expansion. Use entity names as keywords to crawl openly licensed images from Google Images/Bing Images, dynamically filtering out low-quality images (e.g., resolution below 640×480 or heavy text overlay). For example, collect conference site photos and map visualizations for the entity "UN Climate Change Conference," prioritizing higher-relevance images. Concurrently, build an image metadata database recording source URLs and crawl timestamps, saving all images to a unified folder.

Perform feature encoding on multimodal data: Resize images to 256×256 resolution, apply center-cropping to 224×224, and use histogram equalization to mitigate illumination bias. Load ResNet-152 with ImageNet pre-trained weights, remove the original fully connected layer, and replace it with an identity function to directly output a 2048-dimensional feature vector after global average pooling. (e.g., encode conflict scene images into feature patterns highlighting military equipment and structural damage).

Process image feature tensors obtained via ResNet-152. Aggregate historical event texts, relations, and timestamps by entity ID. Use Python for file processing to consolidate all data into the ICEWS+ folder for subsequent research.

Through these steps(see Fig. 1), a comprehensive multimodal temporal knowledge graph EA dataset (ICEWS+) will be constructed, exhibiting enhanced diversity and authenticity.

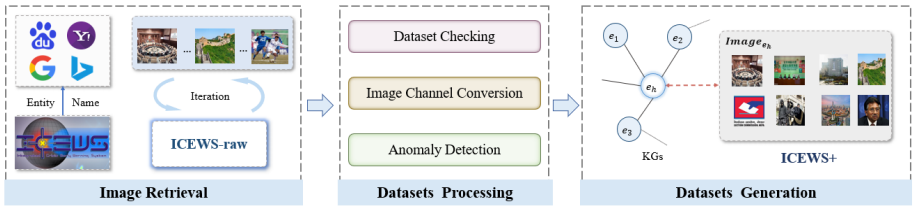

Fig. 1. Dataset Construction Process

3.2 Details

ICEWS+ is a multimodal temporal knowledge graph EA dataset proposed by us. It is the first such dataset both domestically and internationally, featuring strong authenticity, high information integration, and rich modal diversity. The dataset was developed to address the data requirements for knowledge graph EA in real-world scenarios, providing richer and more authentic data support to advance related research.

The roles of the components are detailed below: **Entity IDs (ent_ids)**: Core entities for alignment, derived from ICEWS data (e.g., political figures, events, countries). Numerical IDs connect to other components. **Reference Entity Pairs (ref_pairs)**: Represents 7,566 aligned entities for use as a validation set in algorithms. **Relation IDs (rel_ids)**: Defines relationships (e.g., "chaired a meeting" or "reached an agreement") to form entity network graphs. **Supervised Entity Pairs (sup_pairs)**: Represents 1,000 entity correspondences for supervised learning training pairs. **Time IDs (time_id)**: Unique timestamps record event times to assist alignment tasks. **Tuples (triples)**: Quadruples that log specific events and their timing (head entity, relation, tail entity, time). **Unsupervised Links (unsup_link)**: Forms the training set (8,168 pairs) in unsupervised mode, with the validation set combining training and validation data. **Image Information (icews_images)**: Images (e.g., of events or figures) in ResNet-processed pickle format, linked to entities via IDs for multimodal integration.

4 The Proposed DynEA

This section proposes the **DynEA** framework, designed to effectively utilize cross-modal contrastive learning and attention mechanisms for modality fusion, addressing the task of multimodal temporal knowledge graph EA.

4.1 Definition

The multimodal temporal knowledge graph EA task processes two distinct knowledge graphs $G_s = (E_s, R_s, T, Q_s, I_s)$ and $G_t = (E_t, R_t, T, Q_t, I_t)$, where E_s and E_t represent entity sets, R_s and R_t denote relation sets, T is the shared time interval set, $Q_s \subseteq E_s \times R_s \times E_s \times T$ and $Q_t \subseteq E_t \times R_t \times E_t \times T$ are temporal quadruple sets, and I_s, I_t correspond to image sets.

The alignment objective is to discover the set of semantically equivalent entity pairs:

$$\phi = \{(e_s, e_t) \in E_s \times E_t \mid e_s \equiv e_t\}, \tag{1}$$

where $\equiv$ denotes semantic equivalence between entities. This mapping ϕ constitutes the solution to the EA problem.

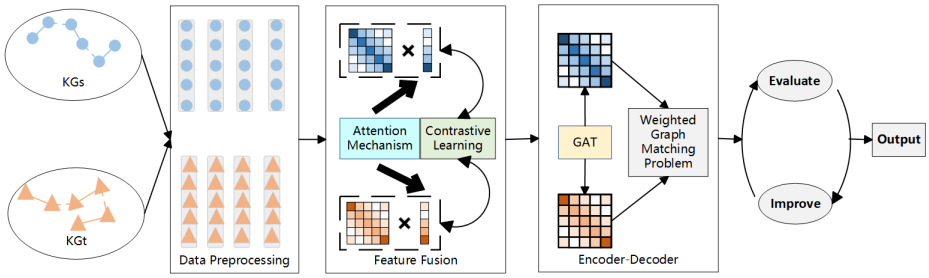

Fig. 2. Overall framework of proposed model

4.2 Multimodal Temporal Data Preprocessing Summary

Knowledge graph construction faces dual challenges: knowledge representation and knowledge acquisition. To address this, we shift to mining unstructured text via NLP for automated entity or relation extraction.

The preprocessing module unifies structured or unstructured data using core techniques—matrix processing, feature processing, data partitioning, and similarity computation—to transform inputs into tensor formats for neural network learning. It employs sparse storage where adjacency matrices per modality are stored in LIL (List of Lists) format using nested lists: an outer list represents rows, while inner lists store non-zero values (`data`) and column indices (`rows`). This structure enables efficient dynamic element addition, rapid row retrieval, and flexible initialization from dense matrices or other sparse formats. This enables efficient handling of multimodal temporal KG datasets [30].

Subsequently, symmetric normalization (analogous to GCN operations) is applied to the adjacency matrix:

$$\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}, \tag{2}$$

where $\tilde{D} = D + I$ and $\tilde{A} = A + I$. This ensures numerical stability in subsequent graph convolutions, preserves information integrity in the adjacency matrix, balances degree heterogeneity, prevents feature distribution shift, maintains graph symmetry, and enhances compatibility with neural networks.

4.3 Feature Fusion

The contrastive learning-based multimodal fusion weight training module extracts structural, relational, attribute, and visual modal features through graph attention networks (GAT) and pre-trained models respectively. It employs intra-modal contrastive loss (ICL) to enhance single-modal discriminability while designing inter-modal alignment loss (IAL) to constrain cross-modal distribution consistency via bidirectional KL divergence. The module dynamically learns modal weights through multi-head self-attention mechanisms and automatically balances multi-objective loss weights using task-dependent uncertainty. Finally, EA is driven by contrastive similarity in a joint embedding space, achieving semantic coordination and adaptive fusion of heterogeneous modal features, which significantly improves EA accuracy in multimodal knowledge graphs.

Intra-modal contrastive loss (ICL) is the core component for modeling entity distinguishability within single modalities in the fusion weight training module.For each modality m, the similarity between entities u and v is defined as the normalized dot product:

$$\delta_m(u, v) = exp(\frac{f_m(u)^\top f_m(v)}{\tau_1}), \tag{3}$$

where $f_m(\cdot)$ is the encoder for modality m (e.g., GAT for structural modality, ResNet for visual modality), τ_1 is the temperature parameter controlling distribution smoothness, and input embeddings are L2-normalized. For a positive sample pair (e_1^i, e_2^i), its alignment probability distribution under modality m is:

$$q_m\left(e_1^i, e_2^i\right) = \frac{\delta_m\left(e_1^i, e_2^i\right)}{\delta_m\left(e_1^i, e_2^i\right) + \sum_{e_1^j \in N_1^i} \delta_m\left(e_1^i, e_1^j\right) + \sum_{e_2^j \in N_2^i} \delta_m\left(e_1^i, e_2^j\right)}. \tag{4}$$

This distribution represents the probability of aligning e_1^i and e_2^i in modality m, with the denominator summing similarities across all positive and negative samples. Since alignment is bidirectional ($e_1^i \leftrightarrow e_2^i$), the reverse probability $q_m(e_2^i, e_1^i)$ is simultaneously computed and averaged to enhance symmetry:

$$\mathcal{L}_m^{\mathrm{ICL}} = -\mathbb{E}_{i \in \mathcal{B}} \log \left[\frac{1}{2}\left(q_m\left(e_1^i, e_2^i\right) + q_m\left(e_2^i, e_1^i\right)\right)\right]. \tag{5}$$

The expectation operation averages over batch data. This cross-entropy-based loss forces alignment probabilities of positive pairs toward 1, effectively enhancing each modality's ability to distinguish aligned entities through bidirectional probability constraints and independent multimodal optimization.

Inter-modal Alignment Loss (IAL) is the core component for modeling cross-modal semantic consistency. The joint embedding is generated by weighted concatenation of unimodal embeddings (structure, name, image, etc.), forming a comprehensive representation that fuses multimodal information. This section treats the prediction distribution of joint embeddings in alignment tasks as the "teacher signal" and unimodal embedding distributions as "student signals". By minimizing their distributional divergence (KL divergence), unimodal embeddings learn cross-modal interactive knowledge, bridging semantic gaps between modalities.

For joint embeddings (modality o, representing fused multimodal information) and unimodal embedding m, the alignment probabilities are defined as:

$$q_o^{'}\left(u,\nu\right)=\frac{\exp(f_o(u)^\top f_o(v)/\tau_2)}{\sum_{k\in\mathbb{N}}\exp(f_o(u)^\top f_o(k)/\tau_2)}, \tag{6}$$

$$q_m^{'}\left(u,\nu\right)=\frac{\exp(f_m(u)^\top f_m(\nu)/\tau_2)}{\sum_{k\in N}\exp(f_m(u)^\top f_m(\nu)/\tau_2)}, \tag{7}$$

where $f_o(\cdot)$ and $f_m(\cdot)$ are encoders for joint and unimodal embeddings respectively, τ_2 is the temperature parameter, and N includes positive/negative samples. IAL measures distribution divergence via bidirectional KL divergence:

$$\mathcal{L}_m^{\mathrm{IAL}}=\mathbb{E}_{i\in\mathcal{B}}\frac{1}{2}\left[\mathrm{KL}\left(q_o^{'}\left(e_1^i,e_2^i\right)\|q_m^{'}\left(e_1^i,e_2^i\right)\right)+\mathrm{KL}\left(q_o^{'}\left(e_2^i,e_1^i\right)\|q_m^{'}\left(e_2^i,e_1^i\right)\right)\right]. \tag{8}$$

Forward KL aligns unimodal embeddings with comprehensive semantics guided by joint embeddings, while reverse KL preserves modal specificity (e.g., visual features in images, text semantics in names). As the core mechanism for cross-modal semantic fusion, IAL injects multimodal interactive knowledge into unimodal embeddings through distribution alignment, eliminating heterogeneity gaps while preserving modal specificity. This design ingeniously incorporates knowledge distillation principles, avoiding noise issues in direct feature fusion and providing an effective cross-modal collaborative solution for multimodal EA.

The attention-based adaptive fusion method concatenates embeddings from four modalities (structure, relation, time, vision) into an input tensor processed by stacked BertLayers. Each BertLayer employs Transformer-based self-attention heads to compute cross-modal attention scores capturing inter-modal dependencies, combined with feed-forward networks for nonlinear enhancement. The tensor propagates through layers, outputting updated features and attention matrices. The final layer's attention matrix is aggregated across heads and row-averaged to produce modality importance weights, which are normalized via Softmax. These weights multiply the original modal embeddings for weighted feature fusion, with the resulting weighted features concatenated into joint embeddings for downstream EA tasks.

4.4 Graph Neural Network-based Encoder-Decoder

The design focuses on cross-domain heterogeneous data fusion and alignment through graph neural networks. In the encoding phase, entities from source and target knowledge graphs (KGs) with mismatched attribute types/quantities are projected into a unified semantic space via shared embedding layers. Temporal and relational encoding handle missing attributes through neighborhood aggregation or adaptive filling strategies. Distribution differences in entity attributes and local topological patterns (e.g., neighbor aggregation paths) are encoded to construct cross-domain alignable structural feature vectors. Temporal and relational features are fused via weighted graph matching to enhance key attributes. Temporal Encoder extracts time-sensitive entity features from quadruples $Q = \{(h, r, t, \tau)\}$:

$$a_{e\tau}^{t} = \frac{\exp\left(|Q_{e\tau}|\right)}{\sum_{\tau' \in T} \exp\left(|Q_{e\tau'}|\right)}, \tag{9}$$

where $|Q_{e\tau}|$ counts quadruples involving entity e in time interval τ. The temporal feature matrix $A^t \in \mathbb{R}^{|E| \times |\mathcal{T}|}$ is split into subject/object matrices and concatenated into $A^t \in \mathbb{R}^{|E| \times 2|\mathcal{T}|}$. Multi-hop graph convolution addresses missing features:

$$H^t = \left[A^t \parallel A \cdot A^t \parallel A^2 \cdot A^t \parallel ... \parallel A^l \cdot A^t\right], \tag{10}$$

yielding $H^t \in \mathbb{R}^{|E| \times 2|\mathcal{T}|(L+1)}$ with multi-hop temporal propagation.

Relation Encoder constructs structural associations with temporal fusion:

$$h_{e_i} = \left[h_{e_i}^{out} \parallel \left(\frac{1}{|\mathcal{N}_e^r|} \sum_{r_j \in \mathcal{N}_e^r} h_{r_j} + \frac{1}{|\mathcal{N}_e^\tau|} \sum_{\tau_j \in \mathcal{N}_e^\tau} h_{\tau_j} \right) \right], \tag{11}$$

initialized as $h_v^{(0)} \sim \text{Glorot}(D)$, where r_{uv} and τ_{uv} denote relations/time between entities u, v. Output: $H^r \in \mathbb{R}^{|E_s \cup E_t| \times D}$.

Decoder frames alignment as weighted graph matching:

$$P_{init} = \alpha \cdot H_s^t \left(H_t^t\right)^T + H_s^r \left(H_t^r\right)^T. \tag{12}$$

Sinkhorn normalization enforces 1-to-1 constraints:

$$\boldsymbol{P}^{(n+1)} = Normalize_{rows}\left(\boldsymbol{P}^{(n)} \odot e^{\lambda \boldsymbol{P}_{init}}\right), \tag{13}$$

$$\boldsymbol{P}^{(n+1)} = Normalize_{columns}\left(\boldsymbol{P}^{(n)}\right). \tag{14}$$

Graph matching objective combines structural/temporal similarity:

$$D_{G_s G_t} = k^r A_s \hat{P} - \hat{P} A_{t2}^2 + k^t A_s^t - \hat{P} A_{t2}^t. \tag{15}$$

Weights derived from Weisfeiler-Lehman kernel:

$$k^r = \hat{k}_{WL}\left(A_s, A_t\right), \tag{16}$$

$$k^t = \hat{k}_{WL}\left(A_s^t, A_t^t\right), \tag{17}$$

$$k_{WL}^{(h)}\left(G_i, G_j\right) = \sum_{k=0}^{h} \left\langle \varphi\left(G_i^{(k)}\right), \varphi\left(G_j^{(k)}\right) \right\rangle. \tag{18}$$

Let h denote the number of iterations for the Weisfeiler-Lehman (WL) graph kernel, and $\phi(G_i)$ represent the feature mapping obtained from the WL isomorphism test. To mitigate the influence of graph size, normalization is typically performed as follows:

$$\hat{k}_{\mathrm{WL}}\left(G_1, G_2\right) = \frac{k_{\mathrm{WL}}\left(G_1, G_2\right)}{\sqrt{k_{\mathrm{WL}}\left(G_1, G_1\right) \cdot k_{\mathrm{WL}}\left(G_2, G_2\right)}}. \tag{19}$$

Finally, we perform a grid search over the range R_α to find the optimal α that minimizes the graph matching objective function $\mathcal{D}_{G_s G_t}$, thereby obtaining the final alignment matrix.

5 Experiments

We evaluate our method on ICEWS+ as follows.

5.1 Experimental Setup

Evaluation Metrics. The widely-adopted Hits@N (H@N) (N=1, 10) and Mean Reciprocal Rank (MRR) are used as the evaluation metrics.Hits@N (expressed as a percentage) measures the ratio of correctly aligned entities appearing within the top-N ranked positions in the alignment matrix $\hat{P}$. MRR (Mean Reciprocal Rank) computes the average of the reciprocal ranks of the first correctly aligned entities in $\hat{P}$, where the reciprocal rank for an entity is the inverse of its highest correct alignment position. Higher values for both Hits@N and MRR correspond to superior EA accuracy.

Baselines. We utilize seven state-of-the-art EA methods as baselines, ensuring fairness by excluding external edge information and conducting experiments solely on our multimodal temporal knowledge graph dataset. Brief descriptions of each method are:

- **MTransE** [6]: Translation-based multilingual KG embedding with cross-lingual transformations for entity/relation alignment.
- **JAPE** [29]: Joint attribute-preserving embedding combining structural and attribute information in a unified space.
- **AlignE** [28]: Alignment-oriented KG embedding using shared space and uniform sampling for cross-lingual alignment.
- **GCN-Align** [34]: GCN-based method integrating structural and attribute data for unified space embedding.
- **RREA** [22]: GNN-based approach with relation-specific embeddings via reflection transformations for discriminative alignment.

- **TREA** [35]: Temporal-aware entity alignment leveraging GNNs and attention mechanisms for time-sensitive KGs.
- **TEA** [17]: Temporal KG entity alignment fusing temporal and relational signals to enhance accuracy.

5.2 Comparison

Table 1 summarizes the EA performance on ICEWS+.

Table 1. Main experiment results.

model	Hits@1	Hits@10	MRR
MTransE	10.1	24.1	15.0
JAPE	14.4	29.8	19.8
AlignE	50.8	75.1	59.3
GCN-Align	20.4	46.6	29.1
RREA	72.2	88.3	78.0
TREA	91.4	96.6	93.3
TEA	<u>94.71</u>	**97.26**	<u>95.76</u>
DynEA	**95.41**	<u>97.19</u>	**96.12**

For comparative clarity, in Table 1, the highest evaluation metric scores are highlighted in bold, while the second-highest scores are marked with underlining. Analysis of Table 1 reveals that DynEA not only demonstrates strong performance compared to baseline methods but also exhibits robust capabilities. As one of the state-of-the-art temporal knowledge graph EA methods, TEA serves as a key benchmark. When comparing DynEA and TEA on the identical multimodal temporal knowledge graph dataset under equivalent virtual environments and configurations, DynEA shows clear superiority. Specifically, DynEA achieves a 0.9% higher average Hits@1 score and a 0.36% higher MRR than TEA. This performance advantage primarily stems from two innovations in DynEA: 1) Integration of multimodal visual information 2) Intelligent weight selection through contrastive learning losses and attention mechanisms during multimodal temporal fusion. In contrast, TEA lacks visual modality processing and employs static (non-adaptive) fusion weights under comparable conditions.

5.3 Ablation Study

We remove each component of DynEA, and report H@1, H@10, and MRR in Table 2, which shows metric values when DynEA is deprived of entity information, temporal information, relational information, or visual information. An additional ablation test investigates the removal of the contrastive learning-based weight fusion module (CLM). The final row displays the full DynEA model's

metrics, with bold indicating the highest scores and underlined scores marking the second-highest values.

The full DynEA configuration achieves significantly higher Hits@1 than all ablated variants, while maintaining leading performance in Hits@10 and MRR. This demonstrates that removing any modality degrades model performance, confirming the importance of each information type. The CLM module proves crucial for optimal operation.

Collectively, these findings validate the necessity of multimodal information fusion and fusion strategy selection.

Table 2. Ablation experiment.

state	Hits@1	Hits@10	MRR
Without Entity	93.96	96.22	94.89
Without Time	93.88	96.63	94.97
Without Relation	94.66	96.97	95.61
Without Image	<u>95.00</u>	<u>97.18</u>	<u>95.88</u>
Without CLM	94.44	91.10	95.35
Full Model	**95.41**	**97.19**	**96.12**

5.4 Sensitivity Study

We systematically adjusted the learning rate across values $\{0.1, 0.01, 0.001, 0.0001\}$, with the resulting Hits@1 performance metrics visualized in Figure 3.

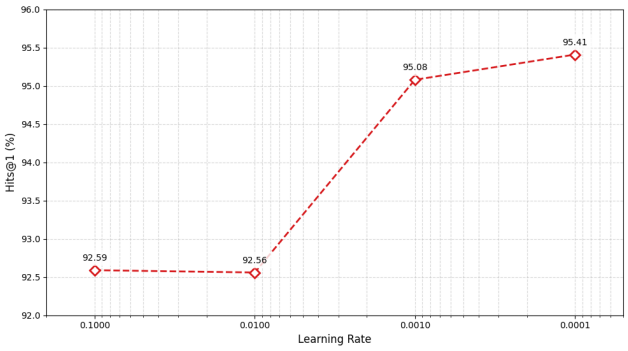

Fig. 3. Impact of learning rate

We observe that as the learning rate decreases, Hits@1 gradually improves. This confirms that higher learning rates tend to cause oscillations that miss opti-

mal solutions, while lower rates may converge to local optima. The dropout rate

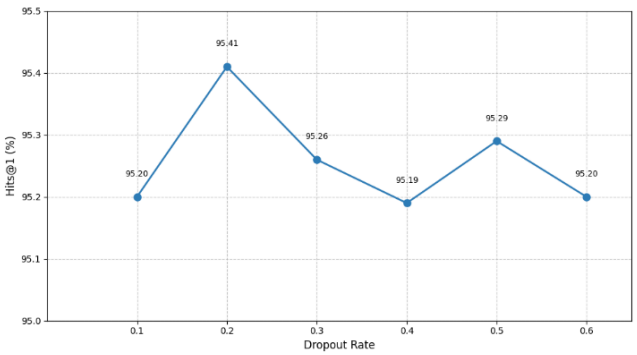

Fig. 4. Impact of dropout rate

was varied within the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, with corresponding Hits@1 results shown in Figure 4.

We observe that variations in dropout rate have minimal impact on Hits@1, yet at $p_{\text{drop}} = 0.2$, Hits@1 is significantly higher. This phenomenon stems from the depth-dependent nature of optimal dropout configuration[33]:

- **Deep networks**: Typically require $p_{\text{drop}} \approx 0.5$ to maximize regularization effects.
- **Shallow architectures**: Should maintain $p_{\text{drop}} < 0.2$ to prevent excessive feature information loss that degrades representational capacity.

Notably, in neural networks of any depth, exceeding the dropout threshold of 0.5 may cause excessive node sparsity. This not only fails to enhance regularization effects but may also disrupt information propagation pathways within the network.

6 Conclusion

This paper addresses data scarcity and fusion inadequacy in multimodal temporal knowledge graph EA by introducing ICEWS+ and proposing the DynEA algorithm. DynEA employs a three-stage framework: multimodal temporal data preprocessing, contrastive learning-based fusion weight training, and GNN encoder-decoder processing. This enables dynamic fusion and precise alignment of cross-modal temporal features. Experiments confirm that the contrastive learning mechanism improves alignment accuracy, while the adaptive weight fusion strategy enhances overall performance, validating the efficacy of multimodal collaboration and dynamic fusion. This work provides novel tools for complex temporal

entity alignment and establishes a foundational dataset, with future research targeting cross-linguistic multimodal alignment and lightweight dynamic fusion to advance knowledge graph alignment toward greater universality and practicality.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., et al.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26 (2013)
2. Bryl, V., Bizer, C.: Learning conflict resolution strategies for cross-language wikipedia data fusion. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 1129–1134. ACM (2014)
3. Cai, L., Mao, X., Ma, M., et al.: A simple temporal information matching mechanism for entity alignment between temporal knowledge graphs. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 2075–2086 (2022)
4. Chen, L., Li, Z., Wang, Y., et al.: MMEA: Entity Alignment for Multi-modal Knowledge Graph. In: International Conference on Knowledge Science, Engineering and Management, pp. 134–147 (2020)
5. Chen, L., Li, Z., Xu, T., et al.: Multi-modal Siamese Network for Entity Alignment. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 118–126 (2022)
6. Chen, M., Tian, Y., Yang, M., et al.: Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. arXiv preprint (2017)
7. Chen, Y., Li, Y., Wen, M., et al.: Survey on Multi-modal Knowledge Graph Fusion Technology. Computer Engineering and Applications **60**(13), 36–50 (2024)
8. Chen, Z., Wu, Y., Feng, Y., et al.: Integrating manifold knowledge for global entity linking with heterogeneous graphs. Data Intelligence, 4(1):20–40 (2022)
9. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
10. Cui, W., Xiao, Y., Wang, H., et al.: KBQA: Learning question answering over QA corpora and knowledge bases. Proceedings of the VLDB Endowment **10**(5), 565–576 (2017)
11. García-Durán, A., Dumancic, S., Niepert, M., et al.: Learning Sequence Encoders for Temporal Knowledge Graph Completion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4816–4821 (2018)
12. Jiang, X., Xu, C., Shen, Y., et al.: Rethinking GNN-based entity alignment on heterogeneous knowledge graphs: New datasets and a new method. arXiv preprint arXiv:2304.03468 (2023)
13. Joy, T., Shi, Y., Torr, P.H.S., et al.: Learning multimodal VAEs through mutual supervision. arXiv preprint (2022)
14. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint (2017)
15. Leblay, J., Chekol, M.W.: Deriving validity time in knowledge graph. In: Companion Proceedings of the Web Conference 2018, pp. 1771–1776. ACM (2018)
16. Liu, F., Chen, M., Roth, D., et al.: Visual pivoting for (unsupervised) entity alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4257–4266 (2021)
17. Liu, X., Wu, J., Li, T., et al.: Unsupervised Entity Alignment for Temporal Knowledge Graphs. arXiv preprint (2023)

18. Liu, Y., Hua, W., Xin, K., et al.: TEA: Time-aware Entity Alignment in Knowledge Graphs. In: Proceedings of the ACM Web Conference 2023, pp. 2591–2599 (2023)
19. Liu, Y., Li, H., García-Durán, A., et al.: MMKG: Multi-modal Knowledge Graphs. In: European Semantic Web Conference, pp. 459–474 (2019)
20. Lu, J., Zhang, J., Feng, J., et al.: A Survey on Temporal Knowledge Graph Construction. Journal of Computer Science and Exploration **19**(2), 295–315 (2025)
21. Mao, X., Wang, W., Wu, Y., et al.: Boosting the speed of entity alignment 10×: Dual attention matching network with normalized hard sample mining. In: Proceedings of the Web Conference 2021, pp. 821–832 (2021)
22. Mao, X., Wang, W., Xu, H., et al.: Relational reflection entity alignment. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1095–1104 (2020)
23. Peng, L., Song, J., Xiong, L., et al.: Advances in Knowledge Fusion Research in Medical Domain. Computer Engineering and Applications **60**(9), 48–64 (2024)
24. Scarselli, F., Gori, M., Tsoi, A.C., et al.: The graph neural network model. IEEE Transactions on Neural Networks 20(1), 61–80 (2009)
25. Song, X., Bai, L., Liu, R., et al.: Temporal Knowledge Graph Entity Alignment via Representation Learning. In: Database Systems for Advanced Applications, pp. 391–406 (2022)
26. Sun, C., Jin, Y., Shen, D., et al.: Enhancing Knowledge Graph Attention by Temporal Modeling for Entity Alignment with Sparse Seeds. In: Database Systems for Advanced Applications, pp. 639–655 (2023)
27. Sun, Z., Zhang, Q., Hu, W., et al.: A benchmarking study of embedding-based entity alignment for knowledge graphs. Proceedings of the VLDB Endowment, 13(12):2326–2340 (2020)
28. Sun, Z., Hu, W., Zhang, Q., et al.: Bootstrapping entity alignment with knowledge graph embedding. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 4396–4402 (2018)
29. Sun, Z., Hu, W., Li, C.: Cross-lingual entity alignment via joint attribute-preserving embedding. arXiv preprint (2017)
30. Tan, Z.: Research on Construction and Representation Technology of Knowledge Graph for Unstructured Data. Ph.D. Dissertation, National University of Defense Technology (2018)
31. Tang, X., Zhang, J., Chen, B., et al.: BERT-INT: A BERT-based interaction model for knowledge graph alignment. interactions, 100:e1 (2020)
32. Wang, C., Huang, Z., Wan, Y., et al.: FuAlign: Cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs. Information Fusion, 89:41–52 (2023)
33. Wang, K.: Research on Rain Attenuation Mitigation Model for Airborne Satellite Communication Systems and Satellite Communication Resource Allocation Methods. Ph.D. Dissertation, University of Electronic Science and Technology of China (2023)
34. Wang, Z., Lv, Q., Lan, X., et al.: Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 349–357 (2018)
35. Xu, C., Su, F., Xiong, B., et al.: Time-aware entity alignment using temporal relational attention. In: Proceedings of the ACM Web Conference 2022, pp. 788–797 (2022)
36. Xu, C., Su, F., Lehmann, J.: 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, pp. 8999–9010 (2021)

37. Yuncan, D., Pingpeng, Y.: Research on adaptive iterative knowledge graph alignment method based on multi-view embedding. Journal Unknown (Incomplete reference)
38. Zhang, F., Yang, L., Li, J., et al.: A Survey on Entity Alignment. Chinese Journal of Computers **45**(6), 1195–1225 (2022)