ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Knowledge Graph Construction Method under Multimodal Information Fusion

Shenghe Zhang[1,*] *and Yongxu Xie*[2]

[1] School of Data Science, City University of Macau, Macau 999078, China
[2] School of Chemical Engineering and Technology, Tianjin University, Tianjin 300354, China

SUMMARY: *This article proposes a knowledge graph construction method under multimodal information fusion, which enhances text semantic information and improves the accuracy of entity recognition and relationship extraction by introducing feature guidance and multimodal cross attention mechanism. The proposed model adopts a multi-level visual cue mechanism and aligns multimodal feature distributions, effectively bridging the semantic gap between text and images and achieving accurate matching of associated objects between entities and images. In terms of model training, Adam optimizer and linear scheduler are used, with different learning rates for language, common sense, visual, and EICF encoders, and a large number of hyperparameter search experiments are conducted to ensure fair comparison. The experimental results on public datasets such as Amazon, YouTube, and self-built datasets show that the proposed model is significantly better than baseline models such as Seq2Seq, NFM, CKE, KGCN, and MMGCN in evaluation metrics such as AUC, AP, and F1. The experimental results have verified the effectiveness and superiority of the proposed model in multimodal information fusion and knowledge graph construction.*

KEYWORDS: *Multimodal; Information fusion; Knowledge graph; Entity recognition; Relation extraction; Adam optimizer*

1 Introduction

In recent years, the rapid development and widespread adoption of artificial intelligence technology have brought unprecedented opportunities and challenges to various industries. Especially since 2020, China has launched large-scale artificial intelligence practices, and the rise and application of generative artificial intelligence technology at the end of 2022 has sparked in-depth discussions and explorations across the global industry [1, 2]. How to effectively utilize cutting-edge digital technologies such as artificial intelligence, knowledge graphs, big data, and metaverse to empower traditional industry sectors, achieve high-quality development in industry sectors, and promote digital transformation in industry sectors has become an urgent task facing various industries [3].

In 2022, OpenAI released the large-scale conversational agent ChatGPT, which quickly became a symbolic milestone in the popularization of generative artificial intelligence and triggered a new wave of innovation in intelligent applications worldwide [4]. Since then, many research institutions and companies have launched their own foundation models and generative AI systems, such as Baidu's "ERNIE Bot", Alibaba's "Tongyi Qianwen", Zhipu AI's "Zhipu Qingyan" and Fudan University's "MOSS", gradually forming a rich ecosystem of general-

*17312257091@163.com
https://doi.org/10.65102/is20261265
Published: 10 June, 2026

purpose and domain-specific models. Behind these products lie a series of large pre-trained models, including BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), GLM (General Language Model) and CLIP (Contrastive Language–Image Pre-training), which are trained on massive text and image corpora and have significantly advanced both natural language processing and computer vision techniques [5]. These foundation models provide powerful semantic representation and cross-modal alignment capabilities, offering a solid technical basis for downstream multimodal understanding tasks. Among them, multimodal named entity recognition has attracted increasing attention from researchers in recent years. In typical application scenarios such as social media, e-commerce reviews and news comments, image resources are often introduced as auxiliary information to supplement the textual description and help disambiguate entity mentions. In a broad sense, multimodal named entity recognition can be regarded as the transformation and alignment of semantic information between multiple modalities such as text and images under a unified task objective, so that the model can jointly exploit heterogeneous yet complementary evidence. In this article, the multimodal named entity recognition task specifically refers to the fusion of text and image information to identify entities in the textual modality while accurately grounding them to the corresponding visual objects, so that the model not only detects entity boundaries and types but also learns to match entities with associated regions in the image [6].

A knowledge graph is a large-scale semantic network and knowledge base with a directed graph structure, which describes the relationships between entities in the objective world in the form of triplets (entity relation entities), where nodes represent entities and edges represent relationships between entities. Knowledge is modeled through structured techniques, and entities are interconnected through relationships to form a network of knowledge structures. With the continuous emergence of a large amount of knowledge in unstructured and semi-structured forms, multimodal data such as images, videos, and audio have begun to receive widespread attention from researchers. Multimodal data makes the demand for cross modal semantic understanding and knowledge representation more urgent. As an important carrier that carries a large amount of underlying knowledge and supports upper-level intelligent applications, multimodal knowledge graphs have become a research hotspot.

2 Related research

2.1 Multimodal learning

A modality is a fine-grained form of information representation that describes how data about the world is encoded, for example as written language, acoustic signals, images, videos or structured sensor readings. When a research object is observed and described through several such channels at the same time, the overall process becomes multimodal, and multimodal learning aims to exploit the complementary information contained in different modalities [7]. Existing studies have shown that, although data from different modalities exhibit heterogeneous statistical properties and noise patterns, they often describe the same underlying entities or events and therefore share latent semantic correlations [8]. To take advantage of this, multimodal learning methods typically adopt one of three basic paradigms: early fusion that directly concatenates or projects raw features into a joint space, late fusion that combines decisions or high-level representations from separate unimodal models, and joint representation learning that seeks a shared semantic space through coordinated encoders and alignment losses. Across these paradigms, a central challenge is how to align heterogeneous modalities and bridge the semantic gap between them while preserving modality-specific nuances that are useful for downstream tasks. In practical applications such as electronic medical records, clinical text,

imaging reports and diagnostic images jointly depict a patient's condition; in human–computer interaction, speech, facial expressions and gestures collectively convey a user's intent. In such scenarios, multimodal learning must not only establish accurate correspondences between modalities, but also design fusion strategies that can selectively attend to informative signals and suppress irrelevant noise, thereby improving the robustness and interpretability of the final model [9].

2.2 Multimodal information extraction

Multimodal information extraction is a key step in constructing multimodal knowledge graphs. This task enhances the semantic expression of text modalities through visual or audio modalities, in order to more accurately identify entities and relationships between entities in the text. It usually includes two sub tasks: multimodal entity recognition and multimodal relationship extraction. As an interdisciplinary research field that combines multimodal learning and natural language processing, the core task of multimodal information extraction is to construct a heterogeneous modal interaction module. This module needs to map data from different dimensions to the same vector space to narrow the semantic gap, and minimize the noise interference caused by modal differences in the process. In multimodal entity recognition tasks, in order to address the challenges posed by the scarcity of labeled samples in named entity recognition research. [10] proposes a multimodal fusion method based on graph alignment to address the issue of insufficient or missing contextual information in the process of relationship extraction. By utilizing visual cues to supplement text semantics, it effectively extracts entity relationships in social comments. To address the issue of multimodal semantic gap. [11] converts image information into text information as an additional semantic supplement and utilizes meta learning methods to alleviate the interference of image noise on the model. In the research of multimodal relationship extraction. [12] designed a two-stage visual fusion network to achieve multi-level deep semantic interaction between images and texts, and validated the effectiveness of the method on a small sample dataset. [13] proposed a universal multimodal alignment framework to bridge the semantic gap between image and text modalities, further enhancing the entity recognition performance of the model.

From existing research it can be observed that most multimodal information extraction systems still decompose entity recognition and relation extraction into two loosely connected subtasks, which are implemented in a pipeline manner. In such frameworks, the outputs of the named entity recognition module are directly fed into the relation extractor, so any boundary errors, type misclassifications or missed entities in the first stage will inevitably propagate to the second stage, leading to cascading errors and reduced robustness. In addition, many methods only introduce visual features at a relatively late stage of the pipeline, so that the interaction between text and image remains shallow and the model cannot fully exploit cross-modal cues when identifying entities and relations. In order to alleviate these problems, this study designs a multimodal joint extraction model that performs entity recognition and relation prediction within a unified framework, and introduces multi-level visual cue modeling and feature-distribution alignment to strengthen the interaction between modalities. By integrating visual information into the representation learning of entities and relations from the outset, the model can more effectively capture fine-grained semantic associations across modalities and thus improve both multimodal entity recognition and relation extraction, as well as the overall quality of the constructed multimodal knowledge graph.

2.3 Multimodal knowledge graph

Multimodal knowledge graph is a knowledge organization model based on semantic networks and supplemented by multimodal data. From the perspective of expression, multimodal knowledge graph is a semantic network that is more complete and closer to the real world by associating text symbols with related multimodal data such as images, sounds, and videos based on text symbols. Table 1 shows the classic methods of knowledge graph representation learning.

In recent years, research on multimodal knowledge graphs in the field of healthcare has attracted significant attention from academia. [14] embedded drug images as entity attributes into a knowledge graph to construct a Chinese medical multimodal knowledge graph CM3KG, and based on this, implemented a "spiritual medicine" medical question answering system. [15] is based on X-rays CT, Six meta paths were designed for four medical modalities of ultrasound and diagnostic text data, and a multimodal knowledge graph was constructed based on a knowledge representation learning model. However, there are numerous types and complex structures of medical data, and how to fully explore the entity and relationship information within them and establish semantic connections between different modalities is currently an unresolved problem [16]. By constructing a multimodal entity relationship joint extraction model to obtain medical entities and their relationships in diagnostic report text, and representing the extracted structured knowledge in the form of triplets, the multimodal data closely related to the diagnostic report, such as medical images and medical record information, are embedded as entity attributes in the triplets. Finally, the constructed multimodal knowledge graph is stored and visualized for analysis.

Table 1: Knowledge Graph Representation Learning Classic Methods

Classification	Typical methods	Description
Distance based model	TransE	(1) Distance–based knowledge graph embedding models such as TransE map entities and relations into a shared low–dimensional vector space and interpret each relation as a translation operation from the head entity embedding to the tail entity embedding, so that correct triples are encouraged to have small distances while incorrect triples are pushed apart. (2) This class of models is conceptually simple, has relatively few parameters and is easy to train on large–scale knowledge graphs, which makes it a strong baseline for link prediction and a convenient building block for more complex architectures in downstream applications. (3) At the same time, the basic translation assumption has limited expressiveness for 1–to–N, N–to–1 and N–to–N relations, so many variants (e.g., TransH, TransR, TransD and STransE) enrich the relation–specific projection space or scoring functions to better capture complex relational patterns while preserving the computational advantages of distance–based modeling.
	TransH	
	TransR	
	TransD	
	STransE	
	TorusE	
	KG2E	
	TransG	
	TransHRA	
	TransRFT	
Model based on random walk	DeepWalk	(1) Learning to aggregate neighboring embedding representations, generate relationship and entity embeddings, can effectively solve the problem of missing triplet predictions. (2) By exploring different node neighborhoods, nodes can be more accurately represented semantically. (3) Utilize first-order information random walks in the network and learn the vector representation of nodes through a word vector model.
	node2ve	
	InGram	
A model based on semantic matching	SLM	(1) Utilize the hierarchical structure information and fine-grained path information between entities to reduce model error representation. (2) Helps with the association between complex semantics, with a simple model but limited to symmetric relationships. (3) By further expressing attention to semantic features, the problem of aggregation models being unable to accurately characterize entities and semantic relationships can be solved.
	NTN	
	DistMult	
	HEAR	
Model based on graph neural network	GNN	(1) Combining transformation-based methods with path level attention mechanisms to address the issues of model error propagation and insufficient interpretability. (2) Effectively processing graph structured data, helping to eliminate ambiguity and reasoning, can be applied to embedding and representing knowledge in multimodal knowledge graphs. (3) Combining relationship graph attention with hypernode graph and combining improved attention mechanism with improved GCN to solve heterogeneous and multi perspective entity representation problems and assist in multimodal knowledge fusion.
	GCN	
	KR-GCN	
	KA-GNN	
	HRGAT	

3 Theoretical model

3.1 Problem statement

The detailed definitions of knowledge graph and collaborative knowledge graph can be found in [17]. Let user $U=\{u_1,\cdots,u_N\}$ represent the set of all N users in the dataset, point set $V=\{v_1,\cdots v_M\}$ represent the set of all M items in the dataset, each user $u\in U$ has a session sequence $S_u=\left\{S_1^u,\cdots,S_{|S_u|}^u\right\}$, $|S_u|$ represent the number of user sequences, each sequence contains interaction items in order and is represented as $S_l^u=\left\{v_1^{u,l},\cdots,v_{|S_l^u|}^{u,l}\right\}$. The current sequence $S_c^u=\left\{S_n^u\right\}$ contains $t-1$ items in order, represented as $S_c^u=\left\{v_1^u,\cdots v_{t-1}^u\right\}$, with the aim of predicting the next $v_t^{\mu_c,l}$ item in the current sequence [18, 19]. All sequences that occurred before S_n^u constitute a set of historical sequences, denoted as $S_h^u=\left\{S_1^u,\cdots,S_{n-1}^u\right\}$. For the target item $v_t^{\mu_c,l}$ in the current session $S_{l^c}^u$ of the current user μ_c to be predicted, all items that occurred before $v_t^{\mu_c,l}$ before the current session $S_{l^c}^u$ form intra session up and down $e_{l,m}=\left\{v_1^{u_c,l},\cdots,v_{t\text{-}1}^{u_c,l}\right\}$, and the historical session set of other users is represented as $S_h^{ou}=\left\{v_1^{ou},\cdots,v_{t-1}^{ou}\right\}$. Specifically, a probability classifier is trained to predict the conditional probability $p\left(v\middle|e_{l,m},S\right)$ of each candidate item $v\in V$, where $S=\left\{S\subset S_h^u\cap S_c^u\cap S_h^{ou}\middle|u\in U\right\}$ includes the set of all historical and current sessions in training. Finally, select items with a conditional probability of top-K to form a recommendation list, the framework of which is shown in Figure 1.

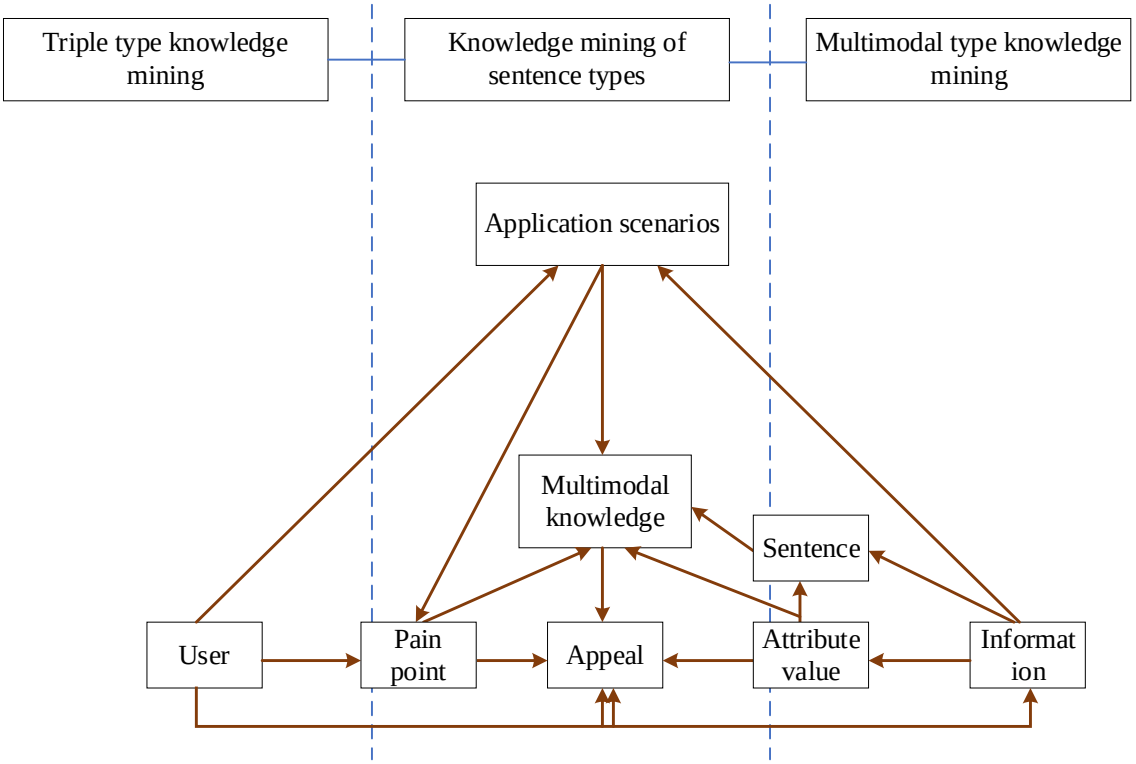


Figure 1: Evolution process of multimodal knowledge mining

Build a historical session graph and a current session graph, and model the complex transformations between projects. All projects appearing in the historical sequence are

represented in a large historical session graph, while the current sequence is represented in a small current session graph [20]. In addition, considering the multiple interests expressed in the sequence, the historical and current conversation graphs are transformed into *k* interest graphs, represented as $G^h = \left(g_1^h, g_2^h, g_3^h, \cdots, g_k^h\right)$ and $G^c = \left(g_1^c, g_2^c, g_3^c, \cdots, g_k^c\right)$ respectively, which capture the complex relationships between similar interest items [21].

3.2 Historical conversation diagram

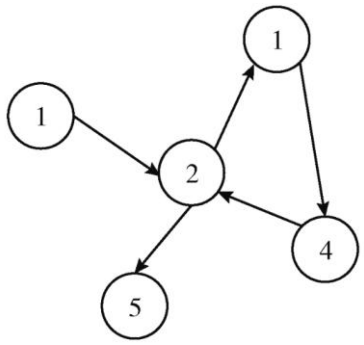
Build a large historical cross session graph G^h for the historical set corresponding to the current sequence, containing items from all historical sequences. $v^h = \left\{v_1^h, \cdots, v_n^h\right\}$ represents the unique set of items represented by nodes, and the connection between two consecutive items in the sequence is represented by directed edges. By inputting the adjacency matrix A_h^i and outputting the adjacency matrix A_h^o, a connection matrix A_h is formed for modeling the complex information propagation between nodes. Figure 2 illustrates an example of constructing input and output adjacency matrices. In the above figure, the same row has been normalized. The connection matrix A_h is shared by the historical interest graph and used as input for GGNN to obtain node embeddings.

Convert all items in the historical sequence into a *D*-dimensional embedding representation $v \in W_e$, where $W_e \in R^{D \times |V|}$ is the embedding matrix of the items, as input to calculate the correlation between the *D*-dimensional embedding representation and *k* items of interest with specific interests.

The above process is expressed as follows:

$$a_{i,m} = v_i^T W_I[:m], m \in \{1, \cdots, k\} \tag{1}$$

where, $a_{i,m}$ is the correlation between project v_i and the *m*-th interest; $W_I[:m]$ is the *m*-th column of W_I.



(a) Knowledge Graph Network

	1	2	3	4	5
1	0	0	0	0	0
2	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	1	0	0	0

(b) Adjacency matrix

Figure 2: Construction of adjacency matrix

To identify the interests that drive users to browse projects, use the calculated relevant scores as input to perform the following operations [22]:

$$y_{i,m} = \frac{\exp\left(\left(\log\left(a_{i,m}\right) + \pi_j\right)/\tau\right)}{\sum_{h-1}^{k} \exp\left(\left(\log\left(a_{i,m}\right) + \pi_h\right)/\tau\right)} \tag{2}$$

where, π_j is the noise with a mean of 0 and a standard deviation of 1 obtained from the distribution; $y_{i,m}$ represents the unique heat vector, indicating that each item is related to a specific interest. In this model, $y_{i,m}$ is set to 0.01.

3.3 Historical interest map

Calculate *k* historical interest maps, denoted as $G^h = \left(g_1^h, g_2^h, \cdots, g_k^h\right)$. In the *m*-th interest map, each node represents a unique term v_i in all historical sequences. Items that do not belong to the *m* interests have no impact on the *m*-th historical interest map and will not appear in it. The interest-based projects related to the *m*-th interest are generated as follows [23]:

$$v_{i,m} = y_{i,m} * v_i \tag{3}$$

$y_{i,m}$ is a scalar close to 0 or 1, depending on whether the user is searching for the *m*-th interest. If $v_{i,m}$ is approximately 0, $v_{i,m}$ is an approximately zero vector in the *m*-th historical interest map, and item v_m will not appear in the *m*-th historical interest map and will not propagate on that interest map. On the contrary, project v_i aggregates information from other projects appearing in the *m*-th historical interest map to prevent insufficient information aggregation between projects associated with different interests. The *m*-th historical interest graph g_m^h is generated by embedding the following initial nodes: $v_m^h = \left[v_{1,m}^k, \cdots, v_{n,m}^k\right]$. In v_m^h, items that do not belong to the *m*-th interest are represented by zero vectors.

4 Construction of multimodal knowledge graph

By analyzing existing multimodal knowledge graphs, fusion techniques can be divided into two categories [24, 25]: (1) fusion techniques for multimodal data. Construct a multimodal knowledge graph by using text transformation, feature extraction, representation learning, and entity alignment techniques for multimodal knowledge. (2) Cross modal knowledge graph fusion technology. Firstly, multi-modal pre training techniques are used to extract features from multi-modal data. Secondly, multi-modal fusion techniques are used to obtain correlations between modalities and fuse their features to construct a multi-modal knowledge graph. The process of constructing a knowledge graph for multimodal data resources is shown in Figure 3.

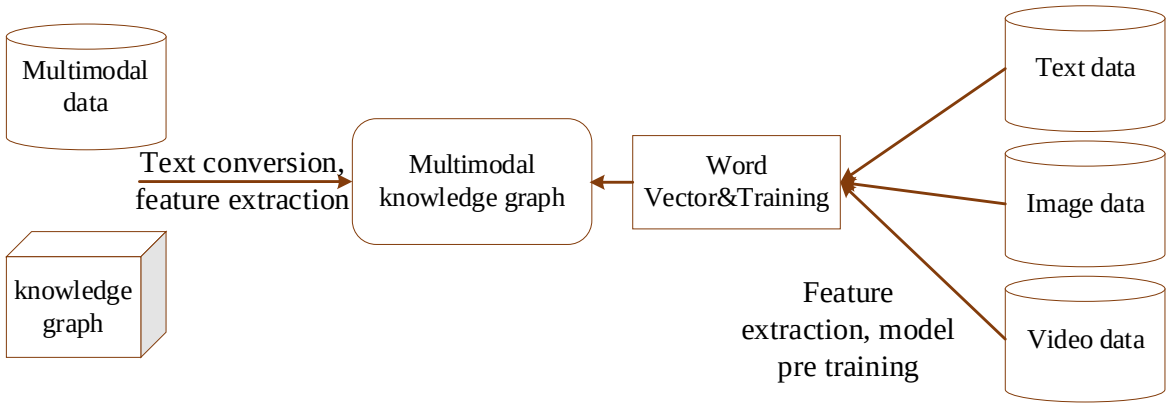


Figure 3: Multi modal knowledge graph fusion technology framework

The specific steps are as follows:

Step 1: Knowledge extraction. In this stage, entities, relations and attributes are extracted from heterogeneous multimodal resources and normalized into structured knowledge elements. For textual data, named entity recognition, relation extraction and keyword mining methods are applied to identify candidate entities and their semantic links, and term-frequency statistics and TF–IDF weighting are used to highlight domain-salient concepts. For images and videos, objects, scenes and other visual concepts are detected and, where appropriate, converted into textual labels so that they can be aligned with the corresponding textual descriptions. Audio or speech data are first transformed into text and then processed by the same natural language processing pipeline. The outputs of this stage provide a unified pool of entity, relation and attribute candidates that will later be fused and mapped into the schema of the knowledge graph.

1) Entity recognition and extraction. In the context of multimodal data resources, entity recognition and extraction mean extracting key information from resource data of different modalities such as text, images, videos, audio, etc. For text modalities, entity extraction mainly relies on natural language processing (NLP) technology for entity recognition of text resources, extracting key concepts, terms, noun phrases, etc. Firstly, it is necessary to remove stop words and perform word segmentation processing. The main purpose of removing stop words is to eliminate those words that contribute little to the meaning of the text but appear frequently. The processing method is to traverse every word in the text through the stop word list, and remove it if it is in the stop word list. The final result is a text that does not contain stop words. Word segmentation is the process of dividing continuous text into independent lexical units. For Chinese text, word segmentation refers to breaking down sentences or paragraphs into meaningful words. The processing method is to use word segmentation tools (such as jieba word segmentation) to segment the text into independent words based on the boundaries and semantic information of the vocabulary. Based on the preprocessing results, use the TF-IDF method to extract entities from text resources. Calculate the inverse text frequency of words in the text, i.e.:

$$\hat{E}_x = \lg \frac{\beta}{\beta_x} \tag{4}$$

where, the inverse text frequency of $\hat{E}_x$ representing words; β represents the number of resource texts; β_x represents the total number of texts containing words in the resource text set.

Smooth E_x, i.e.:

$$\hat{E}_x = E_x + 1 \tag{5}$$

where, $\hat{E}_x$ represents the smooth result of E_x.

Next, calculate the frequency of the word, that is:

$$T_x = \frac{\eta_{x,s}}{N} \tag{6}$$

where, the frequency of word *x* represented by T_x; The frequency of $\eta_{x,s}$ representing the word *x* appearing in text *s*; *N* represents the frequency of occurrence of all keywords in text *s*.

Calculate TF-IDF value:

$$F_x = T_x \cdot \hat{E}_x \tag{7}$$

where, F_x represents the TF-IDF value.

According to F_x, sort the words and select the top M words as keywords for the text resource. For image and video modalities, entity extraction is more complex and typically involves computer vision technology. For example, object detection algorithms can be used to identify objects in images, or image segmentation techniques can be used to extract specific regions of video modality. Motion analysis and temporal information can also be utilized to enhance the accuracy of entity extraction. The entity extraction of audio modalities relies on speech recognition (ASR) and audio analysis techniques. ASR can convert audio into text and then use NLP technology for entity extraction.

2) Relationship extraction: Relationship extraction aims to extract the association information between entities from data, and then transform it into a structured knowledge representation for easy storage, querying, and application. Relationship extraction is typically accomplished through a classifier that takes representations of entities and their surrounding text as input and outputs the categories of relationships. Here, deep learning methods are used to capture deep semantic information in text, enabling accurate relationship classification. Through the process of relation extraction, a triplet<entity 1, relation, entity 2>can be obtained. The specific process is as follows. Firstly, the system will identify potential entity pairs from massive text data, which may be people's names, place names, institution names, etc. Subsequently, deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), or more advanced Transformer structures are utilized to perform deep semantic analysis on the context containing these entity pairs. These models can capture complex semantic features and potential relationships between entities in text through multi-layer nonlinear transformations. During the training phase, the model learns how to map the input text representation to predefined relationship categories, such as "place of birth" and "founder". Ultimately, during the testing or application phase, the model will output a series of relationship triplets, each of which clearly indicates a specific relationship between two entities, such as<Steve Jobs, Founder, Apple Inc.>.

3) Attribute extraction: Define attributes for each entity, describing its characteristics and relationships with other classes. For example, the 'Text Resources' class may have attributes such as' Title', 'Author', and 'Content'. For multimodal resources, specific attributes such as "image size," "video duration," "resolution," and "color distribution" may need to be defined. For attributes, deep learning models can also be used for extraction, such as sequence annotation models or question answering systems, to parse text or multimedia data and automatically identify and fill in the values of these attributes. For example, for the title of an article, the system may use Named Entity Recognition (NER) technology to locate and extract it; For the "duration" attribute of video resources, it may be necessary to directly obtain it through video processing algorithms. This process not only requires the model to have strong text understanding and analysis capabilities, but also needs to combine domain specific knowledge and techniques to ensure the accuracy and comprehensiveness of attribute extraction. Ultimately, each entity will be assigned a detailed description containing multiple attribute value pairs, which provides rich structured data support for subsequent data management, knowledge graph construction, and intelligent applications.

Step 2: Knowledge Fusion: The extracted knowledge elements are fused to eliminate redundancy and contradictions, forming a unified knowledge representation.

Step 3: Create an instance. Instance creation refers to the instantiation of entities. Instance creation can enrich the content of entities, making the concepts and attributes in entities have more specific practical meanings. In the process of instance creation, it is necessary to carefully consider the relationships between instances and add data attributes to the corresponding instances. After adding all instances to the entity, the construction of the data resource

knowledge graph is completed.

Step 4: Storage and Visualization of Knowledge Graph: Knowledge storage is an indispensable part of the knowledge graph construction process, and this study mainly uses the Neo4j graph database to construct the knowledge graph. To store knowledge files in the Neo4j graphics database, you need to first download the corresponding extension neosemantics jar package. The name of the downloaded jar package is neosemantics 4.4.0.3.jar. After the download is complete, create a new database in Neo4j named neo4j (version 4.4.12), open the plugins folder of the neo4j database, copy the knowledge jar package to this folder, and restart the database. After restarting, open the neo4i database management homepage and enter the Cypher statement. After the above steps, the knowledge has been implemented in the storage of Neo4i. In the Neo4i database, by using the built-in visualization module, a knowledge graph of multimodal data resources can be obtained, achieving visualization of the knowledge graph.

5 Experimental analysis

5.1 Dataset

This study uses two widely adopted public heterogeneous graph datasets, Amazon and YouTube, to evaluate the performance and generalization ability of the proposed model under standard benchmark settings [26]. Both datasets contain rich user–item interaction records, textual descriptions and image information, and have been pre-processed into heterogeneous graph structures with predefined training, validation and test splits. To further examine the effectiveness of the method in a domain-specific environment, we additionally construct a proprietary dataset (“ours”) by cleaning and integrating real business object data, which features more complex multimodal attribute information and carefully designed high-quality negative samples to simulate realistic recommendation scenarios [27]. For all three datasets, we randomly divide the data into training, validation and test sets at a ratio of 7:1:2, and ensure that users appearing in the validation and test sets also occur in the training set so that user preferences can be reliably estimated. The detailed statistics of the numbers of users, items, interactions and relations for each dataset are summarized in Table 2.

We use two evaluation criteria. Firstly, following the general approach, we consider relationship prediction as a binary classification task, that is, given two entities and a relationship type, to predict whether this combination holds true. We reported the micro-AUC of Receiver Operating Characteristic (ROC) and Precision Recall Curve (PRC), as well as micro-Fl with a fixed cutoff threshold. Secondly, to evaluate the negative sampling technique used in model training, we reported the filtered mean reciprocal rank (MRR). These ranking metrics evaluate how well the model separates positive relationships from all possible negative relationships, which are generated by changing one entity in the positive relationship. Each test case consists of a positive relationship and all possible negative samples with the same relationship type, as well as at least one node in the positive relationship. MRR represents the overall ranking performance of all cases.

Table 2: Statistical Information of the Dataset

Dataset	Train	Valid	Test
Amazon	10493	1499	2998
Youtube	7349	1050	2100
Ours	28160	4023	8046

5.2 Comparison Method

The proposed method model proposed in this article is compared in performance with the following five baseline methods [28, 29]. (1) CKE: By introducing knowledge graph structures, text, and images to improve the performance of recommendation systems, TransR, stacked denoising autoencoder, and stacked convolutional autoencoder methods are used to concatenate and extract semantic representations. (2) Neurofactorization machine (NFM) is a network model that incorporates neural factors (FM) into a neural network and adds a feature hidden layer. (3) Seq2Seq: By introducing a decoder decoder structure for text learning, this structure is called the Encoder Decoder model, which is a variant of RNN and solves the problem of RNN requiring sequences to be of equal length. (4) KGCN: By mining relevant attributes on the knowledge graph, effectively obtaining the correlation between entities. Using the GCN approach to calculate the neighborhood information of an entity in its neighbor collection, in order to enrich the entity's information. (5) MMGCNL: proposes an end-to-end recommendation framework for multi task feature learning, which enriches the semantic representation of items through alternating learning of entity vectors and item vectors obtained through deep learning.

5.3 Result analysis

Through extensive experiments on publicly available datasets such as Amazon, YouTube, and our own constructed dataset, we compared our proposed method model with the baseline model mentioned above. The experimental results are shown in Tables 3, 4, and 5. Through comparative experiments, it can be seen that the Proposed method model generally maintains a close relationship with the baseline models mentioned above. According to the experimental results, it is shown that the introduction of features centered on evaluation indicators greatly improves the performance of the model and plays a significant role in capturing key semantic information and potential hidden information. We use Adam optimizer and linear scheduler to train the Proposed Method model. We use different learning rates for language, common sense, visual, and EICF encoders. The search space for learning rate is {0.0001, 0.0001, 0.000001, 0.000001). Binary cross extraction is used as the loss function. We conducted experiments using {1, 2, 3, 4, 5, 6, 7, 8} encoder layers and {1, 2, 3, 4, 6} cross attention heads for language, common sense, visual, and EICF encoders. For bimodal cross attention, we conducted experiments with multi-layer attention heads. Use Dropout [0.05, 0.30] (uniform distribution) to standardize the model. For other baseline models, we first conducted experiments using the optimal configurations proposed in their respective papers. In addition, we conducted a large number of hyperparameter search experiments for fair comparison.

According to the experimental results on the Amazon dataset reported in Table 3, the Seq2Seq model, which only models sequence information without explicitly exploiting knowledge graph or multimodal signals, exhibits the weakest performance, with an AUC of 0.4854, an AP of 0.4865 and an F1 score of 0.6325. The NFM model introduces high-order feature interactions and therefore achieves a slight improvement, but its AUC (0.5283), AP (0.5248) and F1 (0.6528) remain at a relatively low level. When side information from the knowledge graph is incorporated, the CKE model obtains more substantial gains, reaching an AUC of 0.6320, an AP of 0.6617 and an F1 of 0.6947, which confirms the benefit of explicit structural knowledge. The KGCN and MMGCN models further integrate graph convolution and multimodal content, and MMGCN in particular delivers competitive performance with an AUC of 0.6912, an AP of 0.7087 and an F1 of 0.7135, highlighting the value of multimodal interaction modeling. Our proposed method achieves the best results among all compared models, with an AUC of 0.7125, an AP of 0.7236 and an F1 of 0.7392, outperforming the

strongest baseline across all metrics. This indicates that the designed feature-guided multimodal fusion and joint modeling of entities and relations enable the model to capture more fine-grained semantic signals and to mine potential latent information on the Amazon dataset more effectively than competing approaches.

Table 3: Experimental Results of Amazon Dataset

Model	AUC	AP	F1
Seq2Seq	0. 4854	0. 4865	0. 6325
NFM	0. 5283	0. 5248	0. 6528
CKE	0. 632	0. 6617	0. 6683
KGCN	0. 6805	0. 6883	0. 6947
MMGCN	0. 6912	0. 7087	0. 7135
Proposed method	0. 7218	0. 7236	0. 7392

Table 4: Experimental results of Youtube dataset

Model	AUC	AP	F1
Seq2Seq	0. 5138	0. 5108	0. 4842
NFM	0. 5102	0. 5052	0. 4841
CKE	0. 6357	0. 6013	0. 6436
KGCN	0. 6980	0. 6986	0. 6428
MMGCN	0. 6883	0. 7014	0. 6983
Proposed method	0. 7325	0. 7165	0. 7390

Table 5: Experimental results of Ours dataset

Model	AUC	AP	F1
Seq2Seq	0. 5127	0. 5093	0. 5275
NFM	0. 6463	0. 6514	0. 6539
CKE	0. 6860	0. 6783	0. 6885
KGCN	0. 6901	0. 6990	0. 7013
MMGCN	0. 7093	0. 6883	0. 7011
Proposed method	0. 7216	0. 7169	0. 7259

According to the experimental results of the Youtube dataset shown in Table 4, the Seq2Seq model and NFM model perform poorly, with AUC values of only 0.5138 and 0.5102, AP values of 0.5108 and 0.5052, and F1 values of 0.4842 and 0.4841, respectively. Their performance is similar and at a relatively low level. The CKE model has been improved, with AUC of 0.6357, AP of 0.6013, and F1 of 0.6436. The performance of KGCN model and MMGCN model is relatively similar, with an AUC of 0.6980 for KGCN, an AP of 0.6986, and an F1 of 0.6428; The AUC of MMGCN is 0.6883, AP is 0.7014, and F1 is 0.6983. And our Proposed method model once again stood out, with AUC reaching 0.7325, AP at 0.7165, and F1 at 0.7390, significantly ahead of the baseline model in all indicators.

According to the experimental results of the Ours dataset shown in Table 5, the AUC of the Seq2Seq model is 0.5127, AP is 0.5093, and F1 is 0.5275, but the performance is still not ideal. The NFM model has shown some improvement, with AUC of 0.6463, AP of 0.6514, and F1 of 0.6539. The CKE model performs well, with AUC of 0.6860, AP of 0.6783, and F1 of 0.6885. The performance of KGCN model and MMGCN model is similar, with an AUC of 0.6901, an AP of 0.6990, and an F1 of 0.7013 for KGCN; The AUC of MMGCN is 0.7093, AP is 0.6883, and F1 is 0.7011. Our proposed method model still maintains its advantages, with AUC of

0.7216, AP of 0.7169, and F1 of 0.7259, outperforming the baseline model in all indicators.

6 Conclusion

This article proposes a knowledge graph construction method under multimodal information fusion, which significantly enhances text semantic information and improves the accuracy of entity recognition and relationship extraction by introducing feature guidance and multimodal cross attention mechanism. The proposed model adopts a multi-level visual cue mechanism and aligns multimodal feature distributions, effectively bridging the semantic gap between text and images and achieving accurate matching of associated objects between entities and images. This innovative method is not only progressiveness in theory, but also shows excellent performance in practical application.

Although the construction and application of multimodal knowledge graphs have made remarkable progress in recent years, there are still several important issues that deserve further investigation. First, the model in this paper mainly considers text and image information; in many real scenarios, however, user behaviour and domain knowledge are also expressed through audio, video and other sensor data, and how to extend the current framework to more modalities while maintaining reliable cross-modal alignment remains an open question. Second, although the use of attention mechanisms enables the model to focus on salient entities and relations across modalities, their design is still relatively simple, and there is room to explore more expressive cross-modal attention and gating strategies that can better balance modality-specific and shared information. Third, the experiments in this article are conducted on public and self-built datasets of limited scale; when deployed in large-scale industrial environments, the continuous expansion of data volume and the growth of model complexity will pose new challenges for training efficiency, online inference latency and system robustness. Future work will therefore consider introducing additional modalities such as audio and video, developing more advanced multimodal attention and fusion modules, and designing distributed training, model compression and incremental updating techniques tailored to large-scale multimodal knowledge graphs, so as to further promote the practical application and long-term evolution of multimodal knowledge graph technology.

References

[1] Venugopal V, Olivetti E. MatKG: An autonomously generated knowledge graph in Material Science[J]. Scientific Data, 2024, 11(1): 217.

[2] Matsumoto N, Moran J, Choi H, et al. KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models[J]. Bioinformatics, 2024, 40(6): 353.

[3] Mitropoulou K, Kokkinos P, Soumplis P, et al. Anomaly detection in cloud computing using knowledge graph embedding and machine learning mechanisms[J]. Journal of grid computing, 2024, 22(1): 6.

[4] Agrawal G, Kumarage T, Alghamdi Z, et al. Can knowledge graphs reduce hallucinations in llms?: A survey[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024: 3947-3960.

[5] Wajid M S, Terashima-Marin H, Najafirad P, et al. DTwin-TEC: An AI-based TEC district digital twin and emulating security events by leveraging knowledge graph[J]. Journal of Open Innovation: Technology, Market, and Complexity, 2024, 10(2): 100297.

[6] Abu-Rasheed H, Weber C, Fathi M. Knowledge graphs as context sources for llm-based explanations of learning recommendations[C]//2024 IEEE Global Engineering Education Conference (EDUCON). IEEE, 2024: 1-5.

[7] Ibrahim N, Aboulela S, Ibrahim A, et al. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges[J]. Discover Artificial Intelligence, 2024, 4(1): 76.

[8] Shokrzadeh Z, Feizi-Derakhshi M R, Balafar M A, et al. Knowledge graph-based recommendation system enhanced by neural collaborative filtering and knowledge graph embedding[J]. Ain Shams Engineering Journal, 2024, 15(1): 102263.

[9] Callahan T J, Tripodi I J, Stefanski A L, et al. An open source knowledge graph ecosystem for the life sciences[J]. Scientific Data, 2024, 11(1): 363.

[10] Venkataramanan R, Tripathy A, Kumar T, et al. Constructing a metadata knowledge graph as an atlas for demystifying AI pipeline optimization[J]. Frontiers in Big Data, 2025, 7: 1476506.

[11] Hussien M M, Melo A N, Ballardini A L, et al. Rag-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models[J]. Expert Systems with Applications, 2025, 265: 125914.

[12] Kosasih E E, Margaroli F, Gelli S, et al. Towards knowledge graph reasoning for supply chain risk management using graph neural networks[J]. International Journal of Production Research, 2024, 62(15): 5596-5612.

[13] Hofmeister M, Brownbridge G, Hillman M, et al. Cross-domain flood risk assessment for smart cities using dynamic knowledge graphs[J]. Sustainable Cities and Society, 2024, 101: 105113.

[14] Remy F, Demuynck K, Demeester T. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights[J]. Journal of the American Medical Informatics Association, 2024, 31(9): 1844-1855.

[15] Hofer M, Obraczka D, Saeedi A, et al. Construction of knowledge graphs: Current state and challenges[J]. Information, 2024, 15(8): 509.

[16] Sequeda J, Allemang D, Jacob B. A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise SQL databases[C]//Proceedings of the 7th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA). 2024: 1-12.

[17] Wajid M S, Terashima-Marin H, Najafirad P, et al. Deep learning and knowledge graph for image/video captioning: A review of datasets, evaluation metrics, and methods[J]. Engineering Reports, 2024, 6(1): e12785.

[18] Mitra R, Dongre A, Dangare P, et al. Knowledge graph driven credit risk assessment for micro, small and medium-sized enterprises[J]. International Journal of Production Research, 2024, 62(12): 4273-4289.

[19] Ain Q U, Chatti M A, Meteng Kamdem P A, et al. Learner modeling and recommendation of learning resources using personal knowledge graphs[C]//Proceedings of the 14th Learning Analytics and Knowledge Conference. 2024: 273-283.

[20] Schembera B, Wübbeling F, Kleikamp H, et al. Towards a knowledge graph for models and algorithms in applied mathematics[C]//Research Conference on Metadata and Semantics Research. Cham: Springer Nature Switzerland, 2024: 95-109.

[21] Kabongo S, D'Souza J, Auer S. ORKG-Leaderboards: a systematic workflow for mining leaderboards as a knowledge graph[J]. International Journal on Digital Libraries, 2024, 25(1): 41-54.

[22] Djenouri Y, Srivastava G, Belhadi A, et al. Intelligent blockchain management for distributed knowledge graphs in IoT 5G environments[J]. Transactions on Emerging Telecommunications Technologies, 2024, 35(4): e4332.

[23] Erickson J S, Santos H, Pinheiro V, et al. LLM experimentation through knowledge graphs: Towards improved management, repeatability, and verification[J]. Journal of Web Semantics, 2025, 85: 100853.

[24] Romano J D, Truong V, Kumar R, et al. The Alzheimer's knowledge base: a knowledge graph for Alzheimer disease research[J]. Journal of Medical Internet Research, 2024, 26: e46777.

[25] Pramanik S, Alabi J, Roy R S, et al. Uniqorn: unified question answering over rdf knowledge graphs and natural language text[J]. Journal of Web Semantics, 2024, 83: 100833.

[26] Barile R, d'Amato C, Fanizzi N. LP-DIXIT: Evaluating Explanations for Link Predictions on Knowledge Graphs using Large Language Models[C]//Proceedings of the ACM on Web Conference 2025. 2025: 4034-4042.

[27] Avdeeva Z K, Gavrilov M S, Lemtyuzhnikova D V, et al. Methods for solving the problem of topic segmentation of texts based on knowledge graphs[J]. Journal of Computer and Systems Sciences International, 2024, 63(4): 642-662.

[28] Cimmino A, García-Castro R. Helio: a framework for implementing the life cycle of knowledge graphs[J]. Semantic Web, 2024, 15(1): 223-249.

[29] Androna C M, Mandilara I, Fotopoulou E, et al. Socio-environmental spatial data reduction and integration in knowledge graphs[C]//2024 Panhellenic Conference on Electronics & Telecommunications (PACET). IEEE, 2024: 1-6.