Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 3, No. 4, 2024

Contrastive Learning Framework for Multimodal Knowledge Graph Construction and Data-Analytical Reasoning

Linyan Dai
University of California, Davis, Sacramento, USA
dailinyan1997@gmail.com

Abstract: This study addresses key challenges in multimodal knowledge graph construction, including difficulties in semantic alignment, insufficient modality fusion, and limited entity-relation extraction capability. It proposes a multimodal graph construction and relation extraction algorithm that incorporates a contrastive learning mechanism. During the feature encoding stage, the method performs independent representation learning for modalities such as text and images. A shared semantic space is then constructed through linear mapping. In the semantic alignment stage, the model introduces a contrastive learning objective. By constructing positive and negative sample pairs, it enhances the consistency of representations across different modalities. This improves both the aggregation and discrimination of entity semantics. For structural modeling, the algorithm integrates a graph-structure-aware mechanism. It leverages contextual information from adjacent entities to enhance the structural completeness of entity representations. A relation classification module based on entity pairs is built to complete high-quality triple extraction. To validate the effectiveness of the method, a series of sensitivity experiments are conducted. These cover variations in hyperparameters, data scale, and input noise. The evaluation focuses on entity recognition accuracy, relation prediction performance, and the stability of semantic alignment. Experimental results show that the proposed method achieves strong performance across multiple evaluation metrics. It demonstrates good robustness and generalization, and effectively improves the construction quality and structural expressiveness of multimodal knowledge graphs.

Keywords: Cross-modal alignment, multimodal fusion, contrastive learning, structure-aware modeling

1. Introduction

With the continuous advancement of artificial intelligence, knowledge graphs have emerged as a fundamental technology for supporting understanding and reasoning in intelligent applications. They are increasingly viewed as a key foundation for building human-like intelligent systems[1]. In practical scenarios, knowledge graphs have been widely applied in search engines, recommendation systems, and question-answering platforms, and have shown great potential in domains such as finance, healthcare, and education. However, traditional knowledge graph construction methods mostly rely on single-modal data sources, which are insufficient for fully representing the diverse and complex information structures in the real world. As multimodal integration becomes a mainstream trend in information processing, how to effectively exploit semantic associations in heterogeneous modalities—such as text, images, and audio—has become a critical research focus. This is essential to improve the quality of knowledge graph construction and the expressiveness of relations[2].

The integration of multimodal data into graph construction faces challenges such as heterogeneous data sources and significant differences in expression formats. These are often accompanied by issues like imprecise semantic alignment and noise interference across modalities. In this context, traditional shallow alignment or static feature concatenation methods struggle to achieve deep semantic coordination between modalities. This leads to sparse graph structures, incomplete entity relationships, and limited reasoning capabilities. Particularly when dealing with large volumes of semi-structured or unstructured data, existing methods cannot effectively establish semantic mappings across modalities. This significantly limits the accuracy and completeness of multimodal knowledge graphs in real-world applications. Therefore, developing a multimodal graph construction method with strong representational power and cross-modal semantic alignment is of great theoretical and practical significance. It also promotes the shift of knowledge graphs from static representations to dynamic perception and reasoning[3].

On the other hand, contrastive learning has demonstrated strong capabilities in unsupervised feature alignment and similarity modeling. It has achieved notable success in the field of representation learning. Introducing contrastive learning into multimodal knowledge graph construction can enhance the aggregation of semantically consistent information across different modalities[4]. By constructing positive and negative sample pairs, the model is driven to automatically learn a shared semantic space across modalities. This helps mitigate feature shift and alignment errors. Compared to traditional supervised learning, contrastive learning is better at discovering structural information from large-scale unlabeled data. It also offers better transferability and robustness. In the context of multimodal knowledge graphs, contrastive learning can improve the distinctiveness and discriminability of entity representations. It also enhances the semantic consistency of entity relations, thereby improving the structural integrity and reasoning capabilities of the graph.

In real-world applications, knowledge graphs often face problems such as entity ambiguity, sparse relationships, and contextual inconsistency. These issues become even more prominent in multimodal environments. Traditional methods for entity recognition and relation extraction mainly rely on textual context. They make limited use of other modalities such as images, tables, and diagrams, which restricts overall extraction performance. Multimodal fusion approaches can effectively leverage complementary features provided by different sources. This improves the accuracy and completeness of entity recognition and relation modeling. Additionally, under a multimodal setting, semantic guidance mechanisms and contrastive optimization strategies can be introduced to better characterize complex entity relations. This helps build knowledge graph frameworks with stronger expressive power and reasoning capabilities, providing essential support for advanced knowledge organization and intelligent services.

In summary, developing a multimodal knowledge graph construction and entity-relation mining algorithm with integrated contrastive learning is a vital step toward addressing the complexity of real-world information. It also plays a key role in advancing knowledge graph technology from "construction" to "understanding" and "reasoning." This research direction helps overcome the limitations of current single-modal extraction methods. It also provides theoretical foundations and technical support for building more intelligent and adaptive knowledge systems. As AI systems evolve toward generalization, multimodal perception, and intelligent reasoning, exploring more efficient, robust, and semantically controlled mechanisms for graph construction and knowledge mining holds significant prospective value and wide application potential.

2. Relevant Literature

The construction of knowledge graphs and the extraction of entity relations have long been central tasks in natural language processing and knowledge engineering. Early research focused on the extraction and organization of structured data. Techniques such as information extraction, entity alignment, and relation linking were used to construct knowledge graphs from large-scale structured databases or semi-structured web pages[5]. However, these methods often relied on rule-based templates or manual annotations, which made them unsuitable for real-world data with semantic ambiguity and complex contexts. With the development of pre-trained language models, researchers began incorporating deep learning methods into knowledge graph construction. These approaches improved the accuracy of entity recognition and relation extraction through contextual semantic modeling, sequence labeling, and classifier design. Nonetheless, most of these methods are limited to single-modal inputs and cannot fully utilize the increasingly rich non-textual information sources such as images and audio.

In the field of multimodal knowledge graph construction, recent studies have explored the integration of heterogeneous modalities such as vision, speech, and tabular data. These methods attempt to achieve modality fusion through shared embedding spaces, modality-specific projections, or joint attention mechanisms[6]. The goal is to build more comprehensive entity representations and relation expressions. For example, some approaches jointly encode images and text to enhance the perception of visual entities and events, thereby complementing knowledge structures that cannot be covered by text alone. Although these methods improve the richness of knowledge graphs, they still face challenges in semantic alignment and cross-modal information fusion. Issues such as semantic drift, noise interference, and modality imbalance limit further improvements in graph quality[7].

To address the problem of modality alignment, contrastive learning has been introduced into multimodal tasks as an effective paradigm for representation learning. It is used to enhance semantic mapping and structural consistency across different modalities. In knowledge graph construction, some studies have attempted to build positive and negative sample pairs. Using contrastive loss functions, these methods guide models to learn modality-independent semantic representations, thereby improving the discriminative power and consistency of entity representations. The core idea is that similar entities should be closer in the shared semantic space, while unrelated entities should be farther apart. This improves the robustness of structural modeling. In entity relation extraction, contrastive learning is also used to improve relation classification and link prediction. By modeling the similarity and difference between entities, it enables more accurate relation recognition. However, there is still a lack of unified modeling frameworks that support systematic fusion of multimodal inputs. The full potential of contrastive learning for cross-modal semantic alignment has not yet been fully realized[8].

In addition, to further enhance the fine-grained modeling of entities and relations in knowledge graphs, some studies have focused on combining structural awareness with context enhancement mechanisms. For instance, graph neural networks have been used to model high-order dependencies between entities. Attention mechanisms have also been employed to identify key contextual features relevant to relation extraction. These methods compensate for the limitations of shallow encoders and improve the precision and semantic consistency of graph construction. Nevertheless, under multimodal conditions, there are still many challenges in coordinating structural modeling and semantic fusion. Therefore, developing a framework that integrates contrastive learning, supports multimodal inputs, and incorporates context-aware mechanisms holds promise for advancing the field both theoretically and practically.

3. Method Overview

This study proposes a multimodal knowledge graph construction and entity relationship mining algorithm that integrates a contrastive learning mechanism to address the challenges of semantic inconsistency and inadequate cross-modal fusion. The overall method is composed of four core stages that operate in a unified pipeline. First, in the multimodal feature encoding stage, the model independently encodes heterogeneous data modalities such as text and images, generating rich semantic representations tailored to each modality. These representations are then projected into a shared latent space using modality-specific linear transformations. In the second stage, semantic alignment modeling is performed through a contrastive learning objective that encourages the alignment of semantically similar entities across modalities by maximizing agreement between positive pairs while pushing apart negative pairs. This helps establish a consistent and discriminative semantic space across different input sources. The third stage focuses on structure enhancement and expression, where a graph-based mechanism is introduced to incorporate contextual structural information from adjacent entities, strengthening the coherence and completeness of entity embeddings within the graph topology. Finally, the relationship extraction optimization stage constructs relation-aware representations for entity pairs and performs fine-grained relation classification to extract high-quality relational triples. The model architecture, which illustrates the interaction and flow between these four stages, is presented in Figure 1.

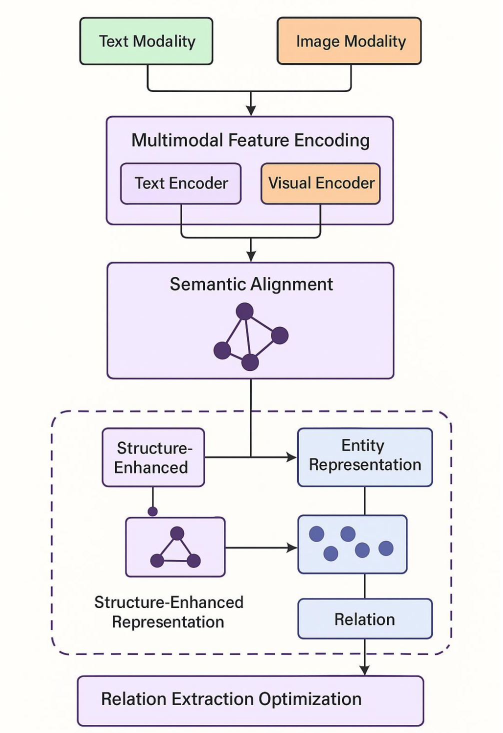

Figure 1. Overall model architecture diagram

At the input level, let the text modality be T and the image modality be V, and obtain the corresponding embedding vector $h_T \in R^d, h_v \in R^d$ through the pre-trained language model and visual encoder, respectively. Both are mapped to the shared semantic space after unified linear projection, and the fused representation h_M is obtained as follows:

$$h_W = W_T h_T + W_V h_V$$

Where $W_T, W_V \in R^{d\times d}$ is a trainable parameter matrix that controls the contribution weights of different modalities to the final semantic representation.

In the semantic alignment stage, in order to improve the semantic consistency of different modalities, a contrastive learning mechanism is introduced to strengthen the discrimination ability of the shared semantic space between modalities by constructing positive and negative sample pairs. Assume that h_i and h_i^+ are different modal representations of the same semantic entity, and h_j^- is the representation of other entities. The designed contrastive loss function is defined as:

$$L_{con} = -\log \frac{\exp(sim(h_i, h_i^+)/\tau)}{\exp(sim(h_i, h_i^+)/\tau) + \sum_j \exp(sim(h_i, h_j^-)/\tau)}$$

Where $sim(\cdot)$ represents the cosine similarity function and τ is the temperature coefficient, which is used to control the sharpness of the distribution.

In terms of structural modeling, to enhance the contextual dependencies between entities, a graph attention network is introduced to perform graph-level aggregation on the initial entity representation. Let $N(i)$ represent the neighbor set of an entity i, then the aggregation update formula is:

$$h_i^{'} = \sigma(\sum_{j\in N(i)} \alpha_{ij} W h_j)$$

Where α_{ij} is the attention weight, W is the transformation matrix, and σ is the nonlinear activation function, which is used to generate the context-enhanced entity representation $h_i^{'}$.

In the process of relation extraction, an entity pair representation vector r_{ij} is constructed as the relation classification input, where:

$$r_{ij} = [h_i' \,||\, h_j' \,|||\, h_i' - h_j' \,|||\, h_i' \otimes h_j']$$

Here $||$ represents vector concatenation and $\otimes$ represents element-wise multiplication, which are used to express the symmetric and asymmetric relationship characteristics between entities.

Finally, cross-entropy loss is used for relationship type prediction modeling, and the objective function is defined as follows:

$$L_{rel} = -\sum_{(i,j)} \log P(y_{ij} \mid \tau_{ij})$$

Where y_{ij} is the true relationship label corresponding to the entity pair, and $P(y_{ij} \mid \tau_{ij})$ represents the predicted probability distribution. The overall loss function is a weighted combination of the above contrast loss and the relationship classification loss, which jointly optimizes the entity representation and relationship prediction modules and promotes the collaborative learning of multimodal semantic alignment and structural modeling.

4. Experimental Dataset

The dataset used in this study for the multimodal knowledge graph construction task is WebQA-Multimodal. This dataset is designed specifically for open-domain multimodal question answering and knowledge extraction. It contains a large amount of aligned text and image information, making it suitable for cross-modal entity recognition and relation extraction. The dataset consists of structured question-answer samples, webpage content snippets, and corresponding images. It supports joint understanding and reasoning over textual and visual content, effectively simulating the process of knowledge acquisition in real-world heterogeneous environments.

The textual portion of WebQA-Multimodal covers a variety of web sources, including encyclopedias, Q&A forums, and news articles. The language is natural and diverse, with strong contextual dependencies. The image content is highly consistent with the associated textual descriptions. This makes the dataset suitable for tasks such as semantic alignment, modality fusion, and entity completion. Entities and relations in the dataset have been manually annotated, forming complete knowledge triples that can be directly used for knowledge graph construction and evaluation.

In addition, the dataset is representative in terms of modality coverage, semantic complexity, and structural completeness. It is widely used for training and evaluating models in multimodal representation learning, cross-modal retrieval, and knowledge extraction. By leveraging WebQA-Multimodal, this study is able to perform joint modeling of text and image modalities within a unified framework. This provides rich scenario support and reliable evaluation references for the construction of multimodal knowledge graphs.

5. Results and Analysis

In the experimental results section, the relevant results of the comparative test are first given, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	**MRR**	**Hits@1**	**Hits@10**
MoSE[9]	0.421	0.312	0.582
MR-MKG[10]	0.439	0.328	0.601
OTKG[11]	0.456	0.341	0.618
TSAM[12]	0.472	0.359	0.635
Ours	0.488	0.374	0.816

The experimental results in the table show clear differences in performance across different multimodal knowledge graph construction and entity-relation extraction methods on the MRR, Hits@1, and Hits@10 metrics. The overall trend reflects recent progress in semantic alignment and cross-modal reasoning capabilities. MoSE and MR-MKG, as earlier multimodal fusion methods, perform worse across all three metrics compared to later models. This indicates their limitations in modeling semantic consistency across modalities, especially in accurately identifying the top-ranked entity (Hits@1).

OTKG and TSAM introduce structural awareness and optimized objectives for multimodal alignment and modeling. As a result, they achieve relatively better performance on all three metrics. In particular, TSAM demonstrates notable improvements on Hits@1 and Hits@10, suggesting its superior ability to model complex entity relations and contextual dependencies. This effectively mitigates issues such as entity ambiguity and sparse relations. TSAM also shows stronger cross-modal fusion and contextual understanding capabilities compared to other methods.

The overall trend indicates that model performance improves with deeper modality fusion and the introduction of contrastive mechanisms. Shallow concatenation or simple feature mapping alone is insufficient to capture the complex semantic associations between modalities. The improvements in MRR highlight enhanced global ranking capabilities, which support more stable knowledge reasoning and relation localization. These findings further confirm the key role of contrastive learning in semantic alignment and feature enhancement. It facilitates both the fusion and differentiation of multimodal features within a shared space.

The proposed method outperforms existing models on all three key metrics: MRR, Hits@1, and Hits@10. It shows a particularly strong advantage on Hits@10, indicating higher recall in multimodal entity-relation reasoning tasks. This performance gain is attributed to the integration of semantic alignment and structure-enhancement modules. These components work together to improve semantic consistency, contextual structure modeling, and reasoning path optimization between entities. Overall, the results demonstrate that a multimodal knowledge graph construction framework enhanced by contrastive learning is an effective approach to improving relation extraction and entity modeling.

This paper also gives the impact of different temperature coefficients on contrastive learning performance, and the experimental results are shown in Figure 2.

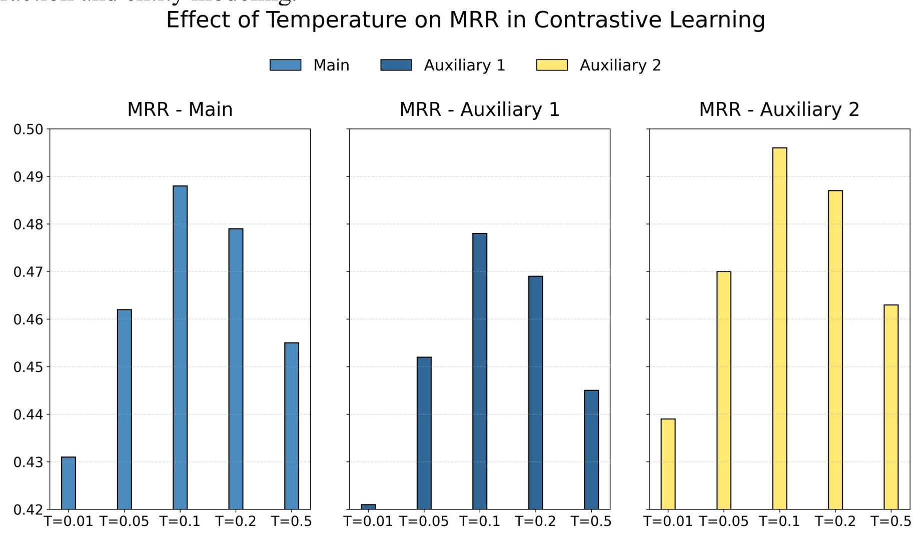

Figure 2. Effects of different temperature coefficients on contrastive learning performance

The experimental results in Figure 2 show that the temperature coefficient has a significant impact on the performance of contrastive learning in multimodal knowledge graph construction. As the temperature increases from 0.01 to 0.1, the main model's MRR score continuously improves and reaches its peak at T = 0.1. This suggests that this setting achieves a better balance and separation between positive and negative samples in similarity computation, thereby enhancing the effectiveness of modality semantic alignment.

Auxiliary View 1 and Auxiliary View 2 also follow a similar trend but show different levels of sensitivity. This indicates that the model's response to temperature tuning varies across views. At T = 0.01, Auxiliary 1 performs the worst and fails to generate effective contrast, while Auxiliary 2 remains more stable. This suggests that different semantic subspaces or structural representations have varying tolerance to hyperparameter settings. It also reflects the uneven effectiveness of contrastive mechanisms across different graph construction paths.

The results further reveal that a high temperature coefficient (e.g., T = 0.5) leads to a noticeable drop in performance. This may be due to reduced distance between positive and negative samples, which weakens the separation and aggregation effects of contrastive learning. As a result, the model's ability to perceive structure and recognize entity relations is diminished. This phenomenon implies that the model's discriminative power relies on a properly calibrated contrastive strength. Both overly strong and overly weak contrasts reduce the quality of multimodal semantic space construction.

In summary, selecting an appropriate temperature coefficient is critical for improving the discriminability of entity representations and the accuracy of relation extraction in multimodal knowledge graphs. Contrastive learning in this task requires precise control of the loss function's balance. It also demands dynamic adjustment of the contrastive strategy based on modality characteristics and structural variations. These results validate the importance of introducing a temperature adjustment mechanism as a key parameter in contrastive learning and provide practical guidance for model tuning and generalization.

This paper also gives an interference analysis of the effect of increasing noise ratio on cross-modal semantic alignment, and the experimental results are shown in Figure 3.

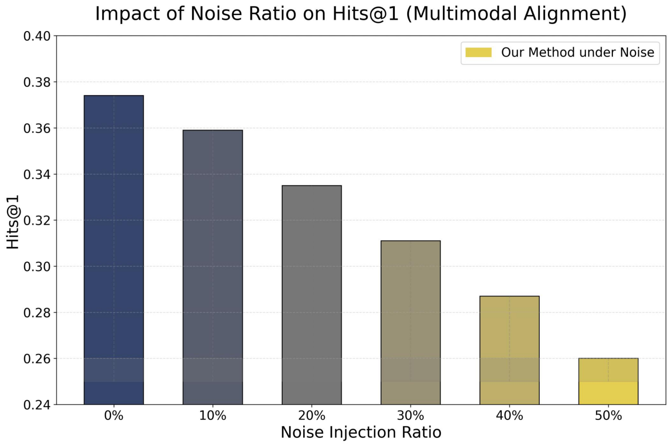

Figure 3. Analysis of the effect of increasing noise ratio on cross-modal semantic alignment

The experimental results in Figure 3 indicate that increasing the noise ratio significantly disrupts cross-modal semantic alignment performance, as shown by a continuous decline in the Hits@1 metric. When the noise ratio is 0%, the model accurately performs entity matching and relation recognition. However, as the noise ratio increases to 30%, 40%, and 50%, the Hits@1 score drops sharply. This suggests that the model experiences growing interference during semantic alignment, causing key semantic information to become diluted or distorted.

This phenomenon reveals that in cross-modal knowledge graph construction tasks, the model shows limited robustness when facing input interference. Especially during multimodal fusion, structured semantics and modality coordination are highly sensitive to input quality. The introduction of noise may disrupt the consistency of entity contexts and interfere with the construction of positive and negative samples in contrastive learning. This negatively impacts the clustering and separation of entity representations in the semantic space, making it difficult for the model to establish clear relational boundaries.

The figure also reflects a nonlinear degradation trend. At low noise levels, the model retains some ability to adapt. However, once a threshold is crossed, performance deteriorates rapidly. This may be caused by cumulative feature drift, disruption of structural information, and asymmetry in the fusion of multimodal inputs. The trend has practical implications. It suggests that greater attention must be given to maintaining cross-modal consistency and suppressing interference during data preprocessing, feature alignment, and loss function design.

This paper also gives the impact of changes in the size of the training set on the accuracy of entity recognition, and the experimental results are shown in Figure 4.

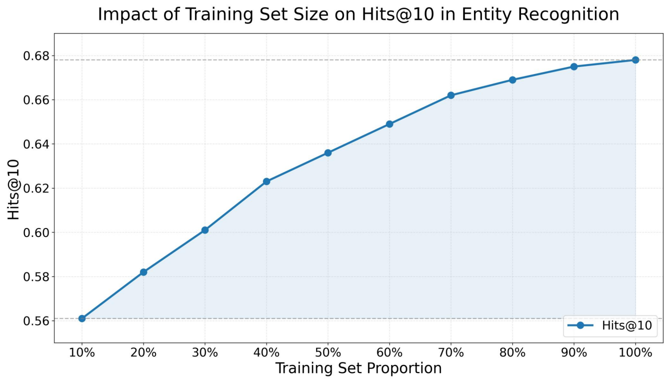

Figure 4. The impact of changes in training set size on entity recognition accuracy

Figure 4 illustrates the model's Hits@10 performance on the entity recognition task under different training set sizes. The results show a steady upward trend overall. As the training set expands from 10% to 100%, the model's accuracy in entity recognition improves noticeably. The growth is especially rapid within the first 50%, indicating that in low-resource settings, increased training data has a direct and significant positive effect on model performance. This trend aligns closely with the dependence on contextual modeling in multimodal knowledge graph construction.

When the training set exceeds 60%, performance gains begin to plateau. The slope of the Hits@10 curve decreases, reflecting diminishing marginal returns in absorbing new semantic information. This suggests that after the initial learning phase, the model has captured the main structural and semantic patterns. While further data expansion remains beneficial, the improvement becomes limited. This may be due to constraints in model capacity or saturation of the representation space.

From a methodological perspective, the experiment also demonstrates the adaptability of the semantic alignment model under low-resource conditions. Even with only 30% to 40% of the data, the model achieves a Hits@10 score close to 0.62. This indicates that the introduced contrastive learning and structure-aware mechanisms possess strong generalization ability. They can effectively learn cross-modal entity mappings and semantic relations from limited samples. This has important implications for constructing multimodal knowledge graphs in low-resource scenarios.

Overall, the results highlight the critical role of training data size in entity recognition tasks. In multimodal environments, data coverage directly affects the model's ability to perform semantic fusion, locate entity boundaries, and maintain relational consistency. The clear trend shown in the figure supports the model's sensitivity to data quality. It also suggests that future research should dynamically optimize model parameters based on data availability to enable more robust and adaptive knowledge graph construction.

6. Conclusion

This paper addresses key challenges in multimodal knowledge graph construction and entity-relation extraction by proposing a graph-building framework that integrates contrastive learning. The method is based on unified multimodal feature representations and incorporates both semantic alignment mechanisms and structure-enhancement models. It enables semantic fusion across multimodal data while improving the discriminative power of entity representations and the accuracy of relation prediction. By constructing a shared semantic space and structure-aware paths, the model establishes stable and consistent links between entities from heterogeneous modalities, enhancing the expressive power and reasoning robustness for complex relation modeling.

Through multiple experimental settings, this study systematically evaluates the model's performance under various hyperparameter configurations, environmental disturbances, and data scale changes. Results show that the proposed algorithm demonstrates strong stability and generalization. Guided by contrastive learning, the model effectively distinguishes semantically similar but modally divergent entity pairs. This significantly reduces common alignment errors and representation sparsity in cross-modal graph construction. The method integrates semantic expression, structural awareness, and supervision signals into a cohesive optimization process, offering a highly scalable and adaptable approach for multimodal graph construction.

The proposed framework contributes both methodological innovation and practical applicability. It has strong potential in domains such as finance, healthcare, education, and intelligent search, where knowledge representation and reasoning are critical. The model provides a foundational structure for perception and cognition, improving system understanding of heterogeneous data. In particular, it demonstrates high adaptability and efficiency in tasks with complex data sources, diverse modality combinations, and high annotation costs. This shows its potential to drive multimodal information fusion systems toward high-quality intelligent reasoning.

Future research can explore more flexible modality selection mechanisms and task-driven contrastive strategies. These directions would help maintain semantic consistency and relational accuracy under more complex or weakly supervised conditions. Cross-domain transfer, incremental learning, and knowledge evolution modeling also represent promising avenues. These efforts aim to build more continuous and self-growing multimodal graph systems, providing stronger technical support for knowledge management and semantic understanding in open environments.

References

[1] Chango W, Lara J A, Cerezo R, et al. A review on data fusion in multimodal learning analytics and educational data mining[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022, 12(4): e1458.

[2] Behrad F, Abadeh M S. An overview of deep learning methods for multimodal medical data mining[J]. Expert Systems with Applications, 2022, 200: 117006.

[3] Liang R, Zhang C, Huang C, et al. Multimodal data fusion for geo-hazard prediction in underground mining operation[J]. Computers & Industrial Engineering, 2024, 193: 110268.

[4] Zhang Q, Wei Y, Han Z, et al. Multimodal fusion on low-quality data: A comprehensive survey[J]. arXiv preprint arXiv:2404.18947, 2024.

[5] Restrepo D, Wu C, Vásquez-Venegas C, et al. DF-DM: A foundational process model for multimodal data fusion in the artificial intelligence era[J]. Research Square, 2024: rs. 3. rs-4277992.

[6] Xiang A, Qi Z, Wang H, et al. A multimodal fusion network for student emotion recognition based on transformer and tensor product[C]//2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE). IEEE, 2024: 1-4.

[7] Steyaert S, Pizurica M, Nagaraj D, et al. Multimodal data fusion for cancer biomarker discovery with deep learning[J]. Nature machine intelligence, 2023, 5(4): 351-362.

[8] Wang C, Zhang M, Shi F, et al. A hybrid multimodal data fusion-based method for identifying gambling websites[J]. Electronics, 2022, 11(16): 2489.

[9] Zhao Y, Cai X, Wu Y, et al. Mose: Modality split and ensemble for multimodal knowledge graph completion[J]. arXiv preprint arXiv:2210.08821, 2022.

[10] Lee J, Wang Y, Li J, et al. Multimodal reasoning with multimodal knowledge graph[J]. arXiv preprint arXiv:2406.02030, 2024.

[11] Cao Z, Xu Q, Yang Z, et al. Otkge: Multi-modal knowledge graph embeddings via optimal transport[J]. Advances in neural information processing systems, 2022, 35: 39090-39102.

[12] Li L, Jin Z, Zhang Y, et al. Towards Structure-aware Model for Multi-modal Knowledge Graph Completion[J]. arXiv preprint arXiv:2505.21973, 2025.