Unifying Within and Across: Intra-Modality Multi-View Fusion and Inter-Modality Alignment for Knowledge Graph Completion

Zhen Li[1], Jibin Wang[1], Zhuo Chen[1], Kun Wu[1], Meng Ai[1], Leike An[1], Liqiang Wang[1], Haoxuan Li[2*]
[1]*China Mobile Information Technology Center*
[2]*Center for Data Science, Peking University*

***Abstract*—Multi-modal knowledge graph completion (MMKGC) enhances the structural and semantic richness of knowledge graphs by integrating diverse information across modalities. However, existing methods often either overlook the diversity within a single modality or fail to ensure effective cross-modality alignment for entity representation. This leads to suboptimal entity representations, as inconsistent or irrelevant data is treated uniformly. To address these challenges, we propose a unified framework that combines intra-modality multi-view fusion with cross-modality alignment (IMVIA for short). Our approach captures the most relevant information within each modality by leveraging relational context. Simultaneously, we apply information disentanglement and contrastive learning, allowing each modality-specific learner to focus on extracting distinctive features while maintaining consistent training objectives across all modalities. Furthermore, we employ a relation-aware gated decision fusion network to robustly integrate diverse information. Experimental results show that IMVIA significantly outperforms state-of-the-art approaches across multiple benchmark datasets, validating its effectiveness and robustness in MMKGC task.**

***Index Terms*—Multimodal, Knowledge Graph Completion, Multi-modal Alignment, Multi-view Fusion, Entity Representation.**

I. Introduction

In recent years, integrating multi-modal data into knowledge graphs (KGs) has become essential for bridging the gap between abstract knowledge and the physical world. Multi-modal knowledge graph completion (MMKGC) addresses the limitations of traditional KGs by incorporating diverse modalities such as text, images, and audio [1], [2]. However, the incompleteness of multi-modal KGs, compounded by the limited availability of multi-modal corpora, continues to hinder their effectiveness. This highlights the need for robust completion methods to fully leverage multi-modal data and improve KG ability of comprehensive representation [3]–[5].

Over the past few years, huge efforts have been made to enhance MMKGC by capturing intra-modality diversity to enrich entity representations [6], [7]. Methods such as VisualBERT [8] and ViLBERT [9] have demonstrated the effectiveness of leveraging variations in textual descriptions and visual perspectives to generate more robust entity embeddings. Recent works, like MNF [10] and MoMoK [11], highlight the importance of capturing nuanced relationships within a single modality through multi-view strategy to enrich the representation space. However, despite these advancements, current approaches still struggle with maintaining consistent objective alignment across different modalities during training, which results in suboptimal entity representations and limits the ability to accurately infer missing links and relationships in multi-modal data.

In parallel, another research direction has focused on cross-modality alignment [12]–[14]. Early works like LXMERT [15] and UNITER [16] focused on creating joint embeddings for visual and textual data, facilitating interaction and alignment between modalities. These foundational approaches motivates the way for more advanced methods, such as ALBEF [17], which introduced contrastive learning techniques to align image and text representations before fusion. Further improvements were presented by models like CLIP [18], [19] and ALIGN [20], which utilized large-scale pre-training for robust image-text matching through contrastive objectives. While these advancements highlight the importance of cross-modality alignment in extracting useful information from different data sources, they do not explicitly encourage each modality-specific learner to focus on learning its own unique features during training, which would further enhance the complementarity of the information learned across modalities [21].

To address these challenges, we propose IMVIA, a unified framework that integrates intra-modality multi-view fusion with cross-modality alignment, effectively leveraging both the diversity inherent within each modality and the complementary information across modalities. Our method begins by capturing intra-modality variability through a multi-view characterization mechanism. Subsequently, we introduce a strategy that combines information disentanglement with contrastive learning, allowing each modality-specific learner to focus on extracting its unique features while maintaining alignment in the training objectives across all modalities. Finally, a relation-aware gated fusion network is introduced to fuse multi-modal decisions while considering the relational context, thereby enhancing the model's reasoning capabilities in complex MMKGC tasks.

II. Proposed Method

In MMKGC, the knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, including entities $\mathcal{E}$ and relations $\mathcal{R}$ with multimodal data. The goal is to infer missing links by learning multimodal embeddings that capture both the structural and semantic characteristics of entities and relations, which are subsequently employed for link prediction in query triples. The proposed framework of IMVIA is illustrated in Fig. 1.

A. Intra-Modality Multi-View Fusion

In MMKGC, entities are often represented through various modalities, such as images, text, or other forms of media. Within each modality, an entity h_i might be associated with multiple views or instances, for example, different images captured under varying conditions or multiple textual descriptions from different sources. The challenge lies in effectively aggregating these multi-view representations within a single modality to form a unified, robust entity representation $\hat{e}_{m,i}$ that accurately captures the characteristic of entity i. In this work, for each modality m (e.g., images, text), we utilize the language-image pre-trained model as described in TinyCLIP [19] to extract features from each view or instance $v_{m,i}$ associated with the entity h_i. Given a set of instances $V_{m,h_i} = \{v_{m,i,1}, v_{m,i,2}, \ldots, v_{m,i,N_i}\}$, the extracted feature vector of the k-th instance in modality m the feature extraction process is represented

*Corresponding Author: hxli@stu.pku.edu.cn

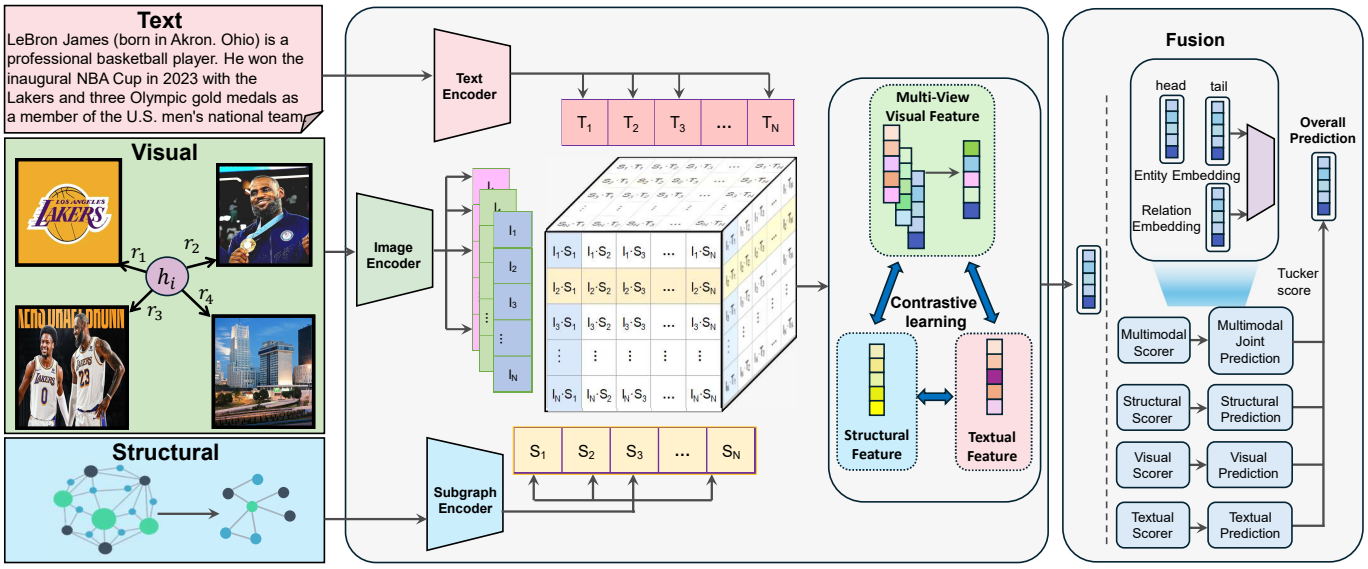

Fig. 1. Overview of the IMVIA framework, consisting of three core components.

as $f_{m,i,k} = \text{CLIP}_m(v_{m,i,k}) \in \mathbb{R}^{d_m}$, where d_m is the dimensionality of the feature space for that modality.

Next, to aggregate these multi-view features within each modality, we introduce a gating mechanism that considers the contribution of each feature based on its relevance to the relational context. The gating vector $g_{m,i,k}$ for each feature $f_{m,i,k}$ is computed as follows:

$$g_{m,i,k} = \sigma(W_g \cdot [e_{h_i}; e_r] + b_g), \tag{1}$$

Here, W_g is a trainable weight matrix, $[e_{h_i}; e_r]$ denotes the concatenation of the entity's structural embedding and the relational embedding, and b_g is a bias term. The function $\sigma(\cdot)$ is a non-linear activation function, such as sigmoid or ReLU. Then the relevance of each feature $f_{m,i,k}$ is modulated by the gating vector, and the importance weight $\alpha_{m,i,k}$ is computed using a similarity function:

$$\alpha_{m,i,k} = \frac{\exp(\text{sim}(g_{m,i,k} \odot f_{m,i,k}, f_{m,i,k}))}{\sum_{j=1}^{N_i} \exp(\text{sim}(g_{m,i,j} \odot f_{m,i,j}, f_{m,i,j}))}, \tag{2}$$

where $\odot$ represents the element-wise product, and $\text{sim}(\cdot,\cdot)$ is a similarity function such as cosine. It ensures that features most aligned with the relational context are given higher importance. Finally, the modality-specific entity representation $\hat{e}_m$ is obtained:

$$\hat{e}_{m,i} = \sum_{k=1}^{N_i} \alpha_{m,i,k} \cdot f_{m,i,k} \tag{3}$$

B. Cross-Modality Disentanglement and Alignment

In MMKGC, it is crucial to ensure that each modality focuses on learning its unique features effectively [11]. To achieve effective disentanglement, we introduce a loss function that minimizes the mutual information between the learned representations of different modalities. This is inspired by the Expert Information Disentanglement strategy [21], which aims to disentangle the learning process across different modalities to maximize their individual contributions. It ensures that each modality specializes in extracting its domain-specific features without interference from other modalities. Thus, for each modality m, the model aims to minimize the overlap of information between the modality-specific embeddings $\hat{e}_{m,i}$ to promote feature disentanglement, which is defined as:

$$L_1 = \frac{1}{|\mathcal{D}|} \sum_{m \neq n} \int_{\hat{e}_{m,i}} \int_{\hat{e}_{n,i}} p(\hat{e}_{m,i}, \hat{e}_{n,i}) \times \log \left(\frac{p(\hat{e}_{m,i}, \hat{e}_{n,i})}{p(\hat{e}_{m,i}) p(\hat{e}_{n,i})} \right) d\hat{e}_{m,i}\, d\hat{e}_{n,i}, \tag{4}$$

where $p(\hat{e}_{m,i}, \hat{e}_{n,i})$ denote the joint probability distribution of the modality-specific embeddings $\hat{e}_{m,i}$ and $\hat{e}_{n,i}$ for entity h_i, while $p(\hat{e}_{m,i})$ and $p(\hat{e}_{n,i})$ are their respective marginal distributions. By minimizing this mutual information, the model encourages each modality encoder to learn specialized features, thereby reducing redundancy across modalities.

While focused modality feature disentanglement is crucial for specialized feature learning, it is equally important to ensure that the learning objectives across different modalities are aligned; otherwise, discrepancies in the training objectives may arise, leading to potential biases across the individual modality feature. Thus, to align the learning objectives across different modalities, we utilize a cross-modality alignment loss L_2, illustrated in Eq. (5), which encourages consistent learning objectives between pairs of modalities, where τ is a learnable temperature parameter that scales the similarity computation. By combining focused modality feature disentanglement with cross-modality alignment loss, IMVIA ensures that each modality contributes its specialized knowledge while maintaining a consistent prediction goal across all modalities.

C. Multi-Modal Joint Decision for Knowledge Graph Completion

In the task of MMKGC, it is essential to leverage all the information provided by textual, visual, structural, and joint modality representations. To achieve this, we introduce a joint representation calculation method that aggregates the specific embeddings from each modality $\hat{e}_{m,i}$ to form a unified entity representation $\hat{e}_{\text{joint},i}$. The joint embedding is computed as:

$$\hat{e}_{\text{joint},i} = \frac{\sum_{m \in \mathcal{M}} \exp\left(W_m^\top P_m(\hat{e}_{m,i})\right) P_m(\hat{e}_{m,i})}{\sum_{m \in \mathcal{M}} \exp\left(W_m^\top P_m(\hat{e}_{m,i})\right)},$$

$$L_2 = \frac{1}{|\mathcal{M}|} \sum_{m \neq n} \left[\mathrm{C}\left(\frac{\exp(\hat{e}_m \cdot \hat{e}_n^T/\tau)}{\sum_{j=1}^{N} \exp(\hat{e}_m \cdot \hat{e}_j^T/\tau)}, \frac{\exp(\hat{e}_n \cdot \hat{e}_m^T/\tau)}{\sum_{j=1}^{N} \exp(\hat{e}_n \cdot \hat{e}_j^T/\tau)} \right) + \mathrm{C}\left(\frac{\exp(\hat{e}_n \cdot \hat{e}_m^T/\tau)}{\sum_{j=1}^{N} \exp(\hat{e}_n \cdot \hat{e}_j^T/\tau)}, \frac{\exp(\hat{e}_m \cdot \hat{e}_n^T/\tau)}{\sum_{j=1}^{N} \exp(\hat{e}_m \cdot \hat{e}_j^T/\tau)} \right) \right] \tag{5}$$

where $P_m(\hat{e}_{m,i})$ represents the projection of the modality-specific embeddings for modality m, and $W_m^\top$ is a learnable attention weight matrix specific to each modality. After obtaining the textual, visual, structural, and joint embeddings, we compute the plausibility of a given triple (h, r, t) by utilizing a Tucker score function $S(h, r, t)$, which integrates these modality-specific embeddings to capture high-order interactions across the knowledge graph:

$$S(h, r, t) = \mathcal{W} \times_1 \hat{e}_m(h) \times_2 \hat{e}_m(r) \times_3 \hat{e}_m(t). \tag{6}$$

Here, $\mathcal{W}$ is the core tensor that captures the high-order interactions among the modality-specific embeddings: $\hat{e}_m(h)$ for the head entity h, $\hat{e}_m(r)$ for the relation r, and $\hat{e}_m(t)$ for the tail entity t. This score $S(h, r, t)$ reflects how well the entities and the relation fit together within the KG across different modalities. Therefore, to optimize the model, we define the following loss function for the MMKGC task:

$$\begin{aligned} L_3 = -\sum_{m \in \mathcal{M}} \Big[& \frac{1}{|\mathcal{T}_+|} \sum_{(h,r,t) \in \mathcal{T}_+} \log \sigma(S_m(h, r, t)) \\ & + \frac{1}{|\mathcal{T}_-|} \sum_{(h,r,t') \in \mathcal{T}_-} \log \sigma(-S_m(h, r, t')) \Big], \end{aligned} \tag{7}$$

where $\mathcal{T}_+$ and $\mathcal{T}_-$ are the set of positive and negative triples, respectively. By minimizing this loss function, the model learns to predict missing links in the knowledge graph, effectively utilizing the information from all available modalities.

The final training objective of our model combines three key components, as shown in Eq. (4), Eq. (5), and Eq. (7): the multi-modality disentanglement loss L_1, the alignment loss L_2, and the KGC loss L_3. These components work together to ensure that the model effectively captures modality-specific features, aligns training objective across modality feature learning, and accurately predicts missing links in the knowledge graph. Let λ_1 and λ_2 be weighting parameters, the combined loss function is defined as:

$$L_{\text{total}} = L_1 + \lambda_1 L_2 + \lambda_2 L_3, \tag{8}$$

III. Experiments

In this section, we thoroughly evaluate the effectiveness of our proposed IMVIA model through experiments on three benchmark datasets. We also perform an ablation study to analyze the importance of each modality (textual, visual, structural, and joint) and demonstrate the superiority of multi-modal fusion in MMKGC tasks.

A. Experimental Settings

Datasets. The experiments were carried out using three publicly available benchmark datasets: DB15K [22], MKG-W [23], and MKG-Y [23]. The DB15K dataset was derived from DBPedia [24], and includes images retrieved from a search engine. The MKG-W and MKG-Y are subsets of Wikidata [25] and YAGO [26] knowledge bases. Table I provides a statistical summary of these datasets.

TABLE I
Statistics of Datasets.

Dataset	Ent	Rel	Train	Valid	Test	Image	Text
MMKB-DB15K	12842	279	79222	9902	9904	12818	9078
MKG-W	15000	169	34196	4276	4274	14463	14123
MKG-Y	15000	28	21310	2665	2663	14244	12305

Evaluation Metrics. To evaluate our approach, we perform a link prediction task [27] on the three datasets. Link prediction is a critical task in KG completion, aimed at predicting missing entities in a query $(h, r, ?)$ or $(?, r, t)$. This task can be divided into head prediction and tail prediction. Following with previous studies, we employ the following rank-based metrics [28] like mean reciprocal rank (MRR) and Hit@K (where $K = 1, 3, 10$) as metrics:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}, \text{Hits@K} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathbb{I}(\text{rank}_i \leq K),$$

where $|Q|$ is the total number of queries, rank_i is the rank position of the correct entity for the i-th query, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the condition inside is true, and 0 otherwise. Here We report Hits@1, Hits@3, Hits@10, and Hits@100 in our experiments.

Baselines. To demonstrate the effectiveness of our approach, we compared it with several existing SOTA methods, which can be grouped into three categories. **Uni-modal KGC methods:** These methods only use the structural information of the KGs to learn embeddings, including: TransE [27], DistMult [29], ComplEx [30], RotatE [28], PairRE [31], GC-OTE [32]. **Multi-modal KGC methods:** These methods leverage both structural and multi-modal information in KGs, including: IKRL [33], TBKGC [34], TransAE [35], MMKRL [36], RSME [37], VBKGC [38], OTKGE [39]. **Negative Sampling methods:** These methods enhance the KGC performance by generating high-quality negative samples, including: KBGAN (TransE) [40], MANS [41], MMRNS [23].

Implemention Details. We conducted experiments on the DB15K dataset using the PyTorch framework, running on a single NVIDIA GeForce RTX4090 GPU. The best-performing hyperparameters were found using grid search on the validation set. The candidate hyperparameter ranges were as follows: batch size of 1024, number of epochs set to 2000, dropout rate of 0.3, learning rate of 0.001, embedding dimension of 256. For more information on additional parameters and model configurations, please refer to our code available at: https://anonymous.4open.science/r/ICASSP2025_IMVIA_Code-14B0/.

B. Performance Comparison

Table II presents the main results of our method compared to these baselines on the three datasets. As shown in this table, our method consistently outperforms the baselines across all datasets. On the MMKB-DB15K dataset, our model achieves a 5.61% improvement in MRR, 7.33% in Hits@1, 4.43% in Hits@3, and 2.54% in Hits@10 compared to the best baseline methods. Similar trends are observed on the MKG-W and MKG-Y datasets, where our method continues to lead across multiple metrics. It is worth noting that our model demonstrates significant improvement in Hits@1, indicating that it is particularly effective at making precise predictions and ranking the correct entity first. This suggests that our method more effectively captures the relational and multi-modal information, leading to more robust and accurate KGC. Additionally, our model shows strong generalization capability across different datasets, further validating its effectiveness in diverse scenarios.

C. Ablation Study

To evaluate the contribution of each modality (textual, visual, structural, and the fused joint representation) in our model, we conducted an ablation study. We systematically removed each modality

TABLE II

THE MAIN MMKGC RESULTS ON THREE DATASETS. THE BEST RESULTS ARE BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED. WE ALSO REPORT THE IMPROVEMENT OF OURS COMPARED TO THE OPTIMAL BASELINE.

Model	MMKB-DB15K				MKG-W				MKG-Y			
	MRR ↑	Hit@1↓	Hit@3↓	Hit@10↓	MRR↑	Hit@1 ↓	Hit@3 ↓	Hit@10↓	MRR↑	Hit@1 ↓	Hit@3 ↓	Hit@10↓
TransE [27]	23.03	14.78	26.28	39.59	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
DistMult [29]	27.48	18.37	31.57	45.37	20.99	15.94	22.28	30.86	25.04	19.33	27.80	35.95
ComplEx [30]	29.28	17.87	36.12	49.66	24.93	19.09	26.69	36.73	28.71	22.26	32.12	40.93
RotatE [28]	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
PairRE [31]	31.12	21.62	35.91	49.30	34.40	28.24	36.71	46.04	32.01	25.53	35.84	43.89
GC-OTE [32]	31.85	22.11	36.52	51.18	33.92	26.55	35.96	46.05	32.95	26.77	36.44	44.08
IKRL [33]	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26
TBKGC [34]	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24	33.99	30.47	35.27	40.07
TransAE [35]	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
MMKRL [36]	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69	36.81	31.66	39.79	45.31
RSME [37]	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43	34.44	31.78	36.07	39.09
VBKGC [38]	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88	37.04	33.76	38.75	42.30
OTKGE [39]	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38
KBGAN(TransE) [40]	25.73	9.91	36.95	51.93	29.47	22.21	34.87	40.64	29.71	22.81	34.88	40.21
MANS [41]	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78	29.03	25.25	31.35	34.49
MMRNS [23]	32.68	23.01	37.86	51.01	35.03	28.59	37.49	**47.47**	35.93	30.53	39.07	45.47
Ours	**38.29**	**30.34**	**42.29**	**53.55**	**35.62**	**30.22**	**37.70**	46.35	**37.97**	**35.15**	**39.56**	**43.47**
Improve	+5.61	+7.33	+4.43	+2.54	+0.59	+1.63	+0.21	-1.12	+2.04	+1.39	+0.49	-1.00

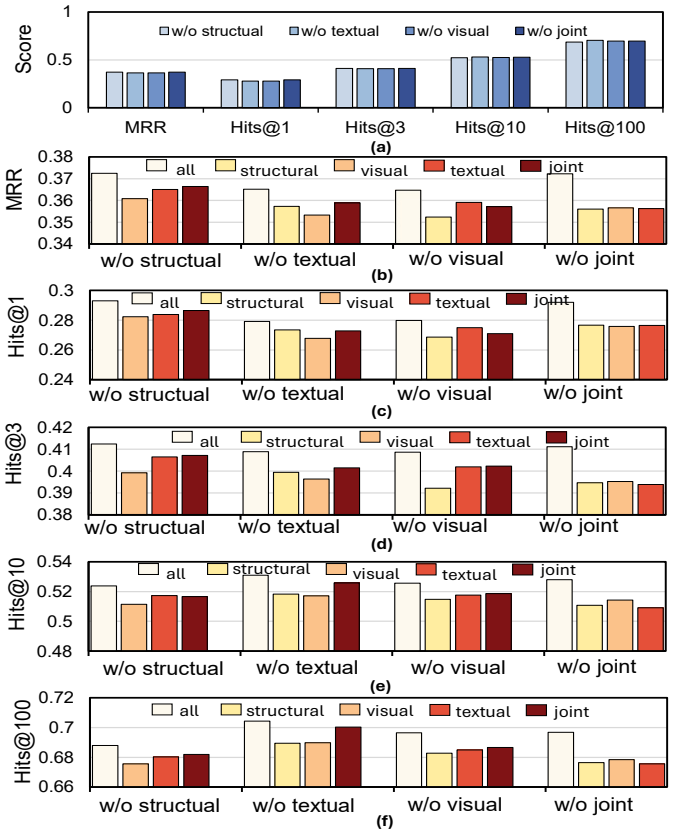

Fig. 2. Performance impact of removing each modality (textual, visual, structural, and joint) on the MMKGC task. The results highlight the importance of the multi-modality fusion process in achieving superior performance.

one at a time and observed the impact on the model's performance. Specifically, we trained four ablated versions of our model: **Without Structural Modality**: Removed the structural features from the multi-modal fusion. **Without Textual Modality**: Removed the textual features from the multi-modal fusion. **Without Visual Modality**: Removed the visual features from the multi-modal fusion. **Without Joint Modality**: Excluded the joint multi-modal fusion, using only the individual modalities without combining them.

The results of this ablation study are dispalyed in Fig. 2, where we plot the model's performance in terms of MRR, Hits@1, Hits@3, Hits@10 and Hits@100. The comparison between the different settings, as shown in Fig. 2(a), indicates that removing any single modality (such as textual, visual, structural, or joint) results in similar performance degradation. However, upon inspection of each individual "w/o" setting, such as w/o structural, we observe that combining multiple modalities consistently outperforms relying on any single modality alone, as seen in Fig. 2(b)-(f). This indicates that IMVIA effectively captures the unique and complementary information from each modality, leading to a significant enhancement in overall performance.

IV. CONCLUSIONS

In this paper, we introduced a unified framework for MMKGC that first employs a multi-view characterization and fusion mechanism to capture the inherent diversity within individual modalities. We then introduce a cross-modality feature disentanglement and alignment strategy to ensure that each modality learner preserves its uniqueness while complementing others under aligned training objectives. Finally, we design a relation-aware gated decision fusion network to effectively integrate multi-modal information, enhancing the model's reasoning capabilities in complex relational contexts. Extensive experiments on multiple benchmark datasets demonstrate that our method significantly outperforms SOTA approaches, confirming its effectiveness in capturing the nuanced relationships within and across modalities. Future research directions include extending this framework to support additional modalities, exploring adaptive view selection mechanisms, and applying the approach to real-time MMKGC scenarios, thereby broadening its applicability and impact in diverse knowledge graph systems.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (623B2002).

REFERENCES

[1] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li, "Retrieval-augmented generation with knowledge graphs for customer service question answering," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2905–2909.

[2] X. Liu, B. Liang, J. Niu, C. Sha, and D. Wu, "Dual-graph co-representation learning for knowledge-graph enhanced recommendation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, and H. Chen, "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 904–915.

[4] A. Saxena, A. Kochsiek, and R. Gemulla, "Sequence-to-sequence knowledge graph completion and question answering," *arXiv preprint arXiv:2203.10321*, 2022.

[5] X. Liu, X. Li, Y. Cao, F. Zhang, X. Jin, and J. Chen, "Mandari: Multi-modal temporal knowledge graph-aware sub-graph embedding for next-poi recommendation," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1529–1534.

[6] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, and Y. Wang, "A robust audio deepfake detection system via multi-view feature," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 131–13 135.

[7] R. Wang, J. Ao, L. Zhou, S. Liu, Z. Wei, T. Ko, Q. Li, and Y. Zhang, "Multi-view self-attention based transformer for speaker recognition," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 6732–6736.

[8] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[9] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[10] Y. Wang, B. Ning, X. Wang, and G. Li, "Multi-hop neighbor fusion enhanced hierarchical transformer for multi-modal knowledge graph completion," *World Wide Web*, vol. 27, no. 5, p. 53, 2024.

[11] Y. Zhang, Z. Chen, L. Guo, Y. Xu, B. Hu, Z. Liu, W. Zhang, and H. Chen, "Mixture of modality knowledge experts for robust multi-modal knowledge graph completion," *arXiv preprint arXiv:2405.16869*, 2024.

[12] Y. Li, J. Chen, Y. Li, Y. Xiang, X. Chen, and H.-T. Zheng, "Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] Z. Wang, Y. Zhao, H. Huang, J. Liu, A. Yin, L. Tang, L. Li, Y. Wang, Z. Zhang, and Z. Zhao, "Connecting multi-modal contrastive representations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22 099–22 114, 2023.

[14] V. Rajan, A. Brutti, and A. Cavallaro, "Robust latent representations via cross-modal translation and alignment," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4315–4319.

[15] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[16] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations.(2019)," *arXiv preprint arXiv:1909.11740*, 2019.

[17] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[19] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. S. Chen, X. Wang *et al.*, "Tinyclip: Clip distillation via affinity mimicking and weight inheritance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 970–21 980.

[20] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[21] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*. PMLR, 2020, pp. 1779–1788.

[22] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio, and D. S. Rosenblum, "Mmkg: multi-modal knowledge graphs," in *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*. Springer, 2019, pp. 459–474.

[23] D. Xu, T. Xu, S. Wu, J. Zhou, and E. Chen, "Relation-enhanced negative sampling for multimodal knowledge graph completion," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 3857–3866.

[24] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic web*, vol. 6, no. 2, pp. 167–195, 2015.

[25] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[26] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 697–706.

[27] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.

[28] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," *arXiv preprint arXiv:1902.10197*, 2019.

[29] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," *arXiv preprint arXiv:1412.6575*, 2014.

[30] T. Trouillon, C. R. Dance, É. Gaussier, J. Welbl, S. Riedel, and G. Bouchard, "Knowledge graph completion via complex tensor factorization," *Journal of Machine Learning Research*, vol. 18, no. 130, pp. 1–38, 2017.

[31] L. Chao, J. He, T. Wang, and W. Chu, "Pairre: Knowledge graph embeddings via paired relation vectors. arxiv 2020," *arXiv preprint arXiv:2011.03798*.

[32] Y. Tang, J. Huang, G. Wang, X. He, and B. Zhou, "Orthogonal relation transforms with graph context modeling for knowledge graph embedding," *arXiv preprint arXiv:1911.04910*, 2019.

[33] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," *arXiv preprint arXiv:1609.07028*, 2016.

[34] H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018, pp. 225–234.

[35] Z. Wang, L. Li, Q. Li, and D. Zeng, "Multimodal data enhanced representation learning for knowledge graphs," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[36] X. Lu, L. Wang, Z. Jiang, S. He, and S. Liu, "Mmkrl: A robust embedding approach for multi-modal knowledge graph representation learning," *Applied Intelligence*, pp. 1–18, 2022.

[37] M. Wang, S. Wang, H. Yang, Z. Zhang, X. Chen, and G. Qi, "Is visual context really helpful for knowledge graph? a representation learning perspective," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2735–2743.

[38] Y. Zhang, Z. Chen, Y. Fang, L. Cheng, Y. Lu, F. Li, W. Zhang, and H. Chen, "Knowledgeable preference alignment for llms in domain-specific question answering," *arXiv preprint arXiv:2311.06503*, 2023.

[39] Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang, "Otkge: Multi-modal knowledge graph embeddings via optimal transport," *Advances in Neural Information Processing Systems*, vol. 35, pp. 39 090–39 102, 2022.

[40] L. Cai and W. Y. Wang, "Kbgan: Adversarial learning for knowledge graph embeddings," *arXiv preprint arXiv:1711.04071*, 2017.

[41] Y. Zhang, M. Chen, and W. Zhang, "Modality-aware negative sampling for multi-modal knowledge graph embedding," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.