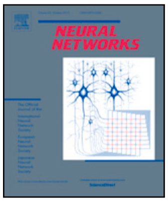
Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

A link prediction method for multi-modal knowledge graphs based on Adaptive Fusion and Modality Information Enhancement

Zenglong Wang, Xuan Liu, Zheng Liu, Yu Weng *, Chaomurilige *

The Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance, Minzu University of China, Beijing, China

ARTICLE INFO

Keywords:
Multimodal knowledge graph completion
Link prediction
Information fusion
Information enhancement

ABSTRACT

Multi-modal knowledge graphs (MMKGs) enrich the semantic expression capabilities of traditional knowledge graphs by incorporating diverse modal information, showcasing immense potential in various knowledge reasoning tasks. However, existing MMKGs encounter numerous challenges in the link prediction task (i.e., knowledge graph completion reasoning), primarily due to the complexity and diversity of modal information and the imbalance in quality. These challenges make the efficient fusion and enhancement of multi-modal information difficult to achieve. Most existing methods adopt simple concatenation or weighted fusion of modal features, but such approaches fail to fully capture the deep semantic interactions between modalities and perform poorly when confronted with modal noise or missing information. To address these issues, this paper proposes a novel framework model—Adaptive Fusion and Modality Information Enhancement(AFME). This framework consists of two parts: the Modal Information Fusion module (MoIFu) and the Modal Information Enhancement module (MoIEn). By introducing a relationship-driven denoising mechanism and a dynamic weight allocation mechanism, the framework achieves efficient adaptive fusion of multi-modal information. It employs a generative adversarial network (GAN) structure to enable global guidance of structural modalities over feature modalities and adopts a multi-layer self-attention mechanism to optimize both intra- and inter-modal features. Finally, it jointly optimizes the losses of the triple prediction task and the adversarial generation task. Experimental results demonstrate that the AFME framework significantly improves multi-modal feature utilization and knowledge reasoning capabilities on multiple benchmark datasets, validating its efficiency and superiority in complex multi-modal scenarios.

1. Introduction

Multi-modal Knowledge Graphs (MMKGs) (Chen, Jia, & Xiang, 2020; Liu et al., 2019) integrate multi-modal information to provide multi-dimensional feature representations for the semantic description of entities, significantly enhancing the semantic expressiveness of knowledge graphs (Liang, Meng et al., 2024; Wang, Mao, Wang and Guo, 2017). In MMKGs, reasoning tasks can effectively uncover potential associations between different modalities, enriching the semantic content of the knowledge graph and improving reasoning capabilities. Among these tasks, link prediction (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013; Jenatton, Roux, Bordes, & Obozinski, 2012; Nayyeri et al., 2021; Rossi, Barbosa, Firmani, Matinata, & Merialdo, 2021; Zhang, Wang, Yang, & Xue, 2022) is one of the critical tasks for reasoning and completing MMKGs. Its goal is to infer missing entity relationships in the knowledge graph to complete incomplete triples. As shown in Fig. 1, different entities in MMKGs are connected by known relationships (solid arrows) and also contain potential relationships yet to be clarified (dotted arrows), such as the causal relationship between the "Great Fire of London" and "Big Ben". The link prediction task aims to infer these missing relationships based on known entity relationships and modality information, thereby improving the completeness of the knowledge graph. However, the link prediction task in MMKGs faces numerous challenges due to the complexity and imbalance of modality information (Chen et al., 2023). First, the quality and representational capacity of different modalities vary significantly; for example, image modalities may lose detail due to low resolution, while textual modalities may suffer from semantic loss due to incomplete descriptions. Furthermore, the interactions between modalities have not been fully explored, and existing methods often fail to capture deep semantic complementarity between modalities. Additionally, multi-modal data often contains noise and missing information, further increasing the difficulty of link prediction. These challenges highlight the urgent need

* Corresponding authors.
E-mail addresses: 23302162@muc.edu.cn (Z. Wang), liuxuan@muc.edu.cn (X. Liu), liuzheng@muc.edu.cn (Z. Liu), wengyu@muc.edu.cn (Y. Weng), chaomurilige@muc.edu.cn (Chaomurilige).

https://doi.org/10.1016/j.neunet.2025.107771
Received 12 January 2025; Received in revised form 11 May 2025; Accepted 16 June 2025
Available online 27 June 2025

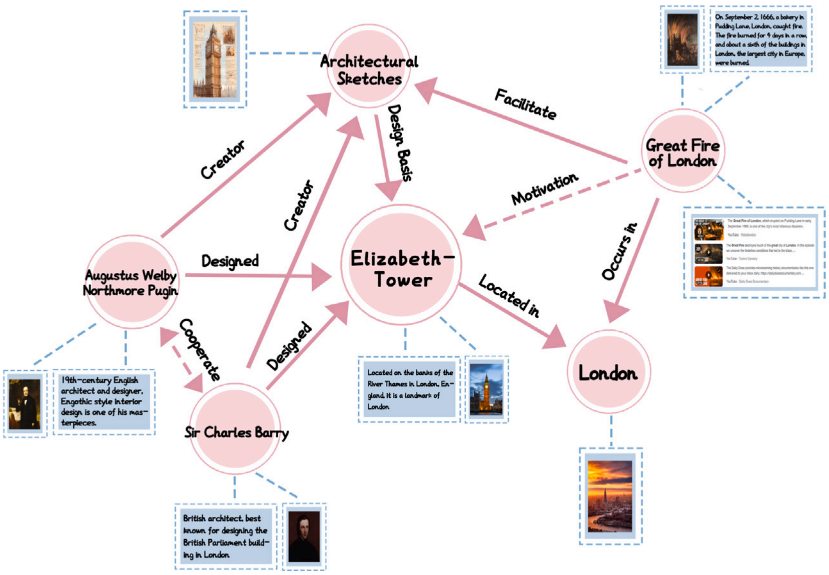

Architectural Sketches
On September 2, 1666, a bakery in Pudding Lane, London, caught fire. The fire burned for 4 days in a row, and about a sixth of the buildings in London, the largest city in Europe, were burned.
Facilitate
Great Fire of London
Design Basis
Creator
Motivation
Creator
Elizabeth-Tower
Augustus Welby Northmore Pugin
Designed
Occurs in
Located in
Cooperate
Designed
London
Located on the banks of the River Thames in London, England. It is a landmark of London
19th-century English architect and designer. Engothic style interior design is one of his masterpieces.
Sir Charles Barry
British architect, best known for designing the British Parliament building in London

Fig. 1. An example of link prediction in a multi-modal knowledge graph. The figure illustrates a multi-modal knowledge graph containing multiple entities and their relationships, where the blue boxes represent the multi-modal information associated with the entities (e.g., text descriptions or images). Solid arrows indicate known entity relationships in the knowledge graph, while dotted arrows represent the entity relationships that need to be inferred and completed (e.g., the causal relationship between the "Great Fire of London" and "Big Ben"). The goal of the link prediction task is to infer these missing relationships based on known entities and modality information, thereby enhancing the completeness of the knowledge graph.

for efficient triple prediction methods and provide clear directions for future research.

To address the challenges posed by the complexity and imbalance of modal information in Multi-modal Knowledge Graph Completion (MMKGC), various methods have been proposed to enhance the performance of triple prediction. For instance, methods based on multi-modal embeddings combine structured information with modal features to effectively improve the semantic representation of entities and relationships in knowledge graphs. A representative work is IKRL (Xie, Liu, Luan, & Sun, 2016), which introduces visual modal features and employs an attention mechanism to dynamically select the modal features most relevant to link prediction, thereby enhancing the accuracy of entity representations. Additionally, MMKRL (Lu, Wang, Jiang, He, & Liu, 2022) jointly optimizes multi-modal embeddings and knowledge reasoning tasks, significantly improving the comprehensive utilization of multi-modal information. In recent years, Generative Adversarial Networks (GANs) have also been applied to link prediction tasks. For example, KBGAN (Wang, Shen et al., 2021) improves the robustness of model training by generating high-quality negative samples, while MMGAT (Chen & Li, 2025) integrates the Graph Attention Mechanism to further explore the interactive relationships among multi-modal information. Although these methods have achieved significant progress in feature modeling and performance improvement, they still face limitations in deeply modeling inter-modal interactions, completing missing modal features, and ensuring robustness against noise interference. As a result, they struggle to fully address the challenges of complex multi-modal scenarios.

Therefore, although existing MMKGC methods have alleviated issues such as the imbalance of modal features and the insufficient utilization of modal information to some extent, numerous challenges remain to be addressed. First, most existing methods rely on shallow inter-modal interaction mechanisms, such as simple concatenation or weighted fusion of modal features (Li, Zhao, Xu, Zhang, & Xing, 2023; Mousselly-Sergieh, Botschen, Gurevych, & Roth, 2018; Wang, Wang et al., 2021), failing to deeply explore the semantic correlations among multi-modal features. Although techniques such as Generative Adversarial Networks (GANs) and self-attention mechanisms have achieved success in various domains, effectively integrating them with complex modality interactions in multi-modal knowledge graph completion remains a challenging problem. Second, there are significant differences in the quality and quantity of different modal features, but current balancing mechanisms struggle to effectively address such modal heterogeneity. This limitation may lead low-quality modal features to negatively impact the overall representation. Moreover, the prevalent noise and missing information in multi-modal data still lack robust solutions, making models susceptible to interference when handling complex real-world scenarios. This hinders the generation of stable and consistent feature representations. Therefore, designing a framework that can deeply explore inter-modal interactions, dynamically balance modal information quality, and enhance the robustness of feature representation has become a critical challenge in further improving the performance of multi-modal knowledge graph link prediction.

To address the aforementioned challenges, this paper proposes a novel multi-modal knowledge graph completion framework—AFME (Adaptive Fusion and Modality Enhancement)—designed to achieve efficient modeling and reasoning optimization of multi-modal information through two key modules. First, the Modal Information Fusion module (MoIFu) introduces a relationship-driven denoising mechanism and a dynamic weight allocation mechanism based on modal confidence and contextual relationships. This significantly improves the representation quality and fusion efficiency of multi-modal features, effectively mitigating issues of imbalance between modalities and interference from noise. Second, the Modal Information Enhancement module (MoIEn), built upon a Generative Adversarial Network (GAN) structure, employs the generator to provide global guidance for other feature modalities based on structural modalities. It also leverages a multi-layer self-attention mechanism to optimize both intra- and inter-modal features. Simultaneously, the discriminator uses a deep network to evaluate the authenticity and consistency of the generated features, further enhancing the model's capability to represent and utilize multi-modal information. This motivates an adaptation of the GAN framework to better suit multi-modal data and the task of knowledge graph completion. The main contributions of this paper can be summarized as follows:

- In the Modal Information Fusion module, a relationship-driven denoising mechanism and a dynamic weight allocation mechanism based on modal confidence and contextual relationships are proposed. These effectively alleviate issues of modality imbalance and information noise.

- In the Modal Information Enhancement module, an optimization scheme based on a Generative Adversarial Network (GAN) is designed. Through the collaborative interaction of the generator and discriminator, the framework enhances the expressive power of multi-modal features and improves the authenticity and consistency of generated modal embeddings.
- The proposed framework is experimentally evaluated on multi-modal datasets such as TIVA and KVC16K, with extensive experiments demonstrating the effectiveness and superiority of AFME in complex multi-modal scenarios.

2. Related work

2.1. Multimodal knowledge graph completion

Knowledge Graph Completion (KGC) (Liang, Meng et al., 2024), as an important research direction in the field of knowledge graphs, aims to infer missing triples in knowledge graphs, thereby enhancing their completeness and reasoning capabilities. Traditional KGC methods (Bordes et al., 2013; Sun, Deng, Nie, & Tang, 2019; Trouillon, Welbl, Riedel, Gaussier, & Bouchard, 2016; Yang, Yih, He, Gao, & Deng, 2014) typically rely on embedding-based models, which embed entities and relationships into continuous vector spaces and evaluate the plausibility of triples using a scoring function (Wang, Mao et al., 2017). Based on different modeling strategies, these methods can be categorized into the following two types: **(1)** Translation-based Models: These methods measure the plausibility of triples by modeling the relationship as a translation from the head entity to the tail entity. For example, the TransE (Bordes et al., 2013) model represents the semantic relationship between entities and relationships through a linear translation approach, while RotatE (Sun et al., 2019) introduces complex vector spaces to support more sophisticated relationship modeling. **(2)** Semantic Matching Models: These methods use tensor decomposition and similarity functions to semantically match the embeddings of entities and relationships. For instance, DistMult (Yang et al., 2014) models the semantic relationships of triples through inner product operations, while ComplEx (Trouillon et al., 2016) extends the semantic representation capabilities with complex-valued embeddings. Some methods also try to extract structural semantics using deep neural networks (Dettmers, Minervini, Stenetorp, & Riedel, 2018; Liang, Liu et al., 2023; Liang, Meng et al., 2023; Yao, Mao, & Luo, 2019; Zhu, Zhang, Xhonneux, & Tang, 2021). However, these approaches rely solely on structured information and fail to fully exploit the multi-modal characteristics of entities. In particular, their performance is significantly limited when dealing with complex multi-modal data.

Multimodal Knowledge Graph Completion (MMKGC) builds upon traditional KGC by introducing multi-modal information (e.g., text, images, audio, and video) to enhance the representational capabilities of knowledge graphs. Current MMKGC methods primarily focus on the following two aspects: **(1)** Modal Fusion and Interaction (Cao et al., 2022; Chen, Fang et al., 2024; Lee, Chung, Lee, Jo, & Whang, 2023; Mousselly-Sergieh et al., 2018; Pezeshkpour, Chen, & Singh, 2018; Wang, Li, Li, & Zeng, 2019; Wang, Wang et al., 2021; Xie et al., 2016): Researchers have designed various modal fusion mechanisms to integrate features from different modalities. For example, OTKGE employs optimal transport theory to dynamically adjust modal weights, enabling the fusion of multi-modal features for more precise feature representation. **(2)** Negative Sampling Enhancement (Xu, Xu, Wu, Zhou, & Chen, 2022; Zhang, Chen, & Zhang, 2023): Multi-modal information is utilized to generate high-quality negative samples to improve the model's training performance. For instance, some studies generate negative samples that closely align with the true distribution by leveraging visual and textual modalities, thereby optimizing the completion performance.

Despite the improvements in MMKGC performance achieved by existing methods, there are still several shortcomings. Many methods focus only on a few modalities (such as text and images) and fail to fully utilize information from various modalities. They lack effective solutions for addressing modal noise and information imbalance. Moreover, inter-modal interaction modeling mostly remains at the level of shallow fusion, failing to fully explore the deep semantic correlations among modalities. To solve these problems, the AFME framework proposed in this paper enhances the efficiency of multi-modal information utilization and improves knowledge reasoning capabilities through a relationship-driven dynamic weight allocation mechanism and a GAN-based enhancement strategy. In recent years, systematic research on MMKGC has continued to advance, and several recent survey studies have provided in-depth overviews of the key challenges in this field. For example, points out that current multi-modal knowledge graph research still lacks effective solutions for addressing modality heterogeneity, semantic inconsistency, and redundant information (Chen, Zhang et al., 2024). In addition, emphasizes that although some progress has been made in modality fusion strategies, there remain pressing challenges in modeling deep semantic interactions and generative completion of missing information (Liang, Meo, De, Tang and Zhu, 2024). To address these issues, the AFME framework proposed in this paper introduces a relationship-driven dynamic weight allocation mechanism to mitigate modality heterogeneity and imbalance. Moreover, it incorporates a GAN-based enhancement strategy that leverages structural modality guidance and self-attention to model deep semantic interactions across modalities and to complete missing modality information. These innovations significantly improve the efficiency of multi-modal information utilization and reasoning capabilities, offering a new solution to core problems in current MMKGC research.

2.2. Generative adversarial network

Generative Adversarial Networks (GANs) are a generative modeling framework consisting of a generator and a discriminator, which optimize each other through a minimax game, enabling the generator to produce samples that approximate the true distribution. Currently, GAN-based research has been widely applied across multiple domains (Croce, Castellucci, & Basili, 2020; Karras, 2019; Wang et al., 2020; Wang, Yu et al., 2017; Wei, Huang, Xia, & Zhang, 2023; Yu, Zhang, Wang, & Yu, 2017; Zhang et al., 2021), with various improvements proposed to enhance its performance. For example, in network optimization research: Wasserstein GAN (WGAN) (Mescheder, Nowozin, & Geiger, 2017)addresses the limitations of the original GAN, such as gradient vanishing and mode collapse, by introducing the Wasserstein distance; WGAN-GP (Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017) further improves training stability by adding a gradient penalty term. In the field of knowledge graph completion, some studies have applied GANs to improve the negative sampling process. For instance, KBGAN (Wang, Shen et al., 2021) combines GAN with reinforcement learning to design an optimized negative sampling strategy, effectively enhancing knowledge graph reasoning performance. Similarly, MMKRL (Lu et al., 2022), targeting multi-modal knowledge graph completion tasks, proposed a GAN-based adversarial training strategy, offering a new solution for complex modality interactions and reasoning.

Although these methods demonstrate certain advantages in specific scenarios, most rely on reinforcement learning modules or overly complex generation strategies, making them less adaptable to more intricate multi-modal scenarios. Moreover, these methods have limited capability to collaboratively optimize the generator and discriminator, resulting in generated features that still lack semantic consistency and authenticity. In practical applications, due to the lack of effective cooperative optimization between the generator and the discriminator, the generated modality features often suffer from semantic drift, content loss, or structural distortion. This not only results in lower overall quality of the generated features, making them insufficient for downstream reasoning tasks that require semantic consistency and authenticity, but may also negatively affect the overall performance of the model in certain cases. To address these limitations, this paper designs a Multi-modal Information Enhancement module (MoIEn) based on Generative Adversarial Networks (GANs) within the AFME framework. The generator leverages the structural modality to guide the generation of other feature modalities and employs a self-attention mechanism to optimize inter-modal interactions. The discriminator, combining multi-layer fully connected networks with a self-attention mechanism, evaluates the authenticity and consistency of the generated features. This design further enhances the modeling capability and reasoning performance of multi-modal information.

3. Task definition

In the multi-modal knowledge graph completion task, the goal is to learn a scoring function F(h,r,t) : $\mathcal{E}\times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$, which is used to evaluate the plausibility of a given triple.

3.1. Multi-modal knowledge graph representation

A multi-ınodal knowledge graph can be defined as a 5-tuple G = ($\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{M}, \mathcal{F}$), where: $\mathcal{E}$ represents the set of entities, and each entity e ∈ $\mathcal{E}$ corresponds to multi-modal feature representations. $\mathcal{R}$ represents the set of relations, and each relation r ∈ $\mathcal{R}$ denotes a semantic relationship between entities. T represents the set of known triples, with each triple denoted as(h,r,t),where h,t ∈ $\mathcal{E}$ represent the head and tail entities, and r ∈ $\mathcal{R}$ represents the relationship $\mathcal{M}$ represents the set of modalities, including text, images, audio, video, and other modalities. $\mathcal{F}$ = $\{f_m^e$ | m∈ $\mathcal{M}$, e∈ $\mathcal{E}\}$ represents the set of modal feature representations, where each entity e has feature representations $\mathrm{f}_{\mathrm{m}}^{\mathrm{e}}$ under modality m. In the multi-modal knowledge graph completion task, the objective is to infer missing triples(h,r,t) based on the known triple set $\mathcal{T}$ and the multi-modal features $\mathcal{F}$, to predict unknown entities h or t, given partial information (e.g., (h,r,?)or(?,r,t)).

3.2. Scoring function

To evaluate the plausibility of a triple (h,r,t), the model needs to define a scoring function ϕ(h,r,t), which calculates the triple score. Plausible triples (positive samples) should have relatively high scores, while implausible triples (negative samples) should have relatively low scores. The scoring function in this paper is modeled by integrating structural information and multi-modal features, and is defined as follows:

$$\phi(h, r, t) = f\left(\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t, \{\mathbf{f}_m^h\}_{m\in\mathcal{M}}, \{\mathbf{f}_m^t\}_{m\in\mathcal{M}}\right) \tag{1}$$

where: e_h and e_t represent the structural embeddings of the head entity h and the tail entity t, respectively. r represents the embedding of the relationship r. $\{\mathrm{f}_\mathrm{m}^h\}_{\mathrm{m}\in\mathcal{M}}$ and $\{\mathrm{f}_\mathrm{m}^\mathrm{t}\}_{\mathrm{m}\in\mathcal{M}}$ represent the embeddings of the head and tail entities in each modality, respectively.

3.3. Negative sampling

During the model training process, negative samples are constructed to assist learning. For each positive sample triple (h,r,t), negative samples (h',r,t) or (h,r,t′) are generated by randomly replacing either the head entity h or the tail entity t. The generation method for negative samples can be defined as follows:

$$\mathcal{T}^- = \{\left(h', r, t\right) | h' \in \mathcal{E}\} \cup \{\left(h, r, t'\right) | t' \in \mathcal{E}\} \tag{2}$$

where $\mathcal{T}^-$ represents the set of generated negative samples. In the inference phase, MMKGC tasks are typically evaluated through link prediction. Given a query triple (e.g.,(h,r,?) or (?,r,t)), the model needs to predict the missing entity. For instance, in a tail entity prediction task, all entities e∈ $\mathcal{E}$ are considered as candidate entities. The score of each candidate triple (h,r,e) is computed, and the candidate entities are ranked based on their scores in descending order. The goal is to rank the true triple as high as possible. Finally, the model performance is typically evaluated using ranking-based metrics, such as Mean Reciprocal Rank (MRR) and Hits@k.

4. Methodology

In this section, we provide a detailed introduction to the proposed multi-modal knowledge graph completion framework. In multi-modal knowledge graphs, differences in modality feature quality, the presence of noise, and the imbalance in information completeness and representation capability across modalities are key factors that affect completion performance. In particular, incomplete modality information and insufficient modeling of inter-modal interactions significantly hinder the effectiveness of reasoning tasks. To address these challenges, this work introduces several core components: a relation-driven denoising mechanism, a dynamic weight allocation mechanism, and a GAN-based structure enhanced with self-attention. The relation-driven denoising and dynamic weighting components dynamically adjust the fusion strategy based on the quality of different modality features and the current contextual relation. The GAN module helps to compensate for missing modality features, while the self-attention mechanism further captures complex global and deep interactions across modalities, thus enhancing the overall feature representation. We name this framework AFME, which addresses the above challenges through two major modules: Modality Information Fusion (MoIFu) and Modality Information Enhancement (MoIEn). Fig. 2 provides an intuitive overview of the model architecture.

4.1. Modality information representation

To fully utilize multi-modal information, the AFME framework first encodes data from each modality to capture deep semantic information and unify their representations. Specifically, we adopt various pre-trained models for feature extraction across different modalities to effectively leverage their heterogeneous characteristics. For the text modality, we utilize the pre-trained language model BERT (Kenton & Toutanova, 2019), which, through large-scale corpus pre-training, effectively extracts both sentence-level and token-level semantic embeddings. For the image modality, we employ classical convolutional neural networks to extract visual features (Karen, 2014), with a focus on information such as object detection and semantic segmentation. For the video and audio modalities, we use established feature extraction methods within their respective domains. These methods encode the spatiotemporal dynamics of videos and the time–frequency characteristics of audio to ensure the semantic completeness and consistency of different modal information. This unified feature encoding provides high-quality inputs for subsequent modal fusion and enhancement processes.

In addition, to unify all modal features into the same vector space, we design a linear projection layer for each modality. This layer maps high-dimensional features into fixed-dimensional vector representations. Meanwhile, for the structural modality embeddings required for knowledge graph triple modeling, we employ trainable parameters within the model training process, rather than relying on pre-trained models. This approach directly represents the structural information of triples in the graph, effectively avoiding constraints from external data and enabling the structural modality embeddings to align more flexibly with the specific task objectives. Specifically, each feature modality f_m in the multi-modal embeddings is represented as:

$$f_m = 1/|X_m(e)| \sum_{i=1}^{|X_m(e)|} \mathsf{PE}_m\left(x_{e,m}^i\right) \tag{3}$$

The text describes that PE_m is the corresponding pre-trained encoder for each modal, and $x_{e,m}^i$ represents the ith modal element in the set $X_m(e)$. Through this approach, the AFME framework captures the

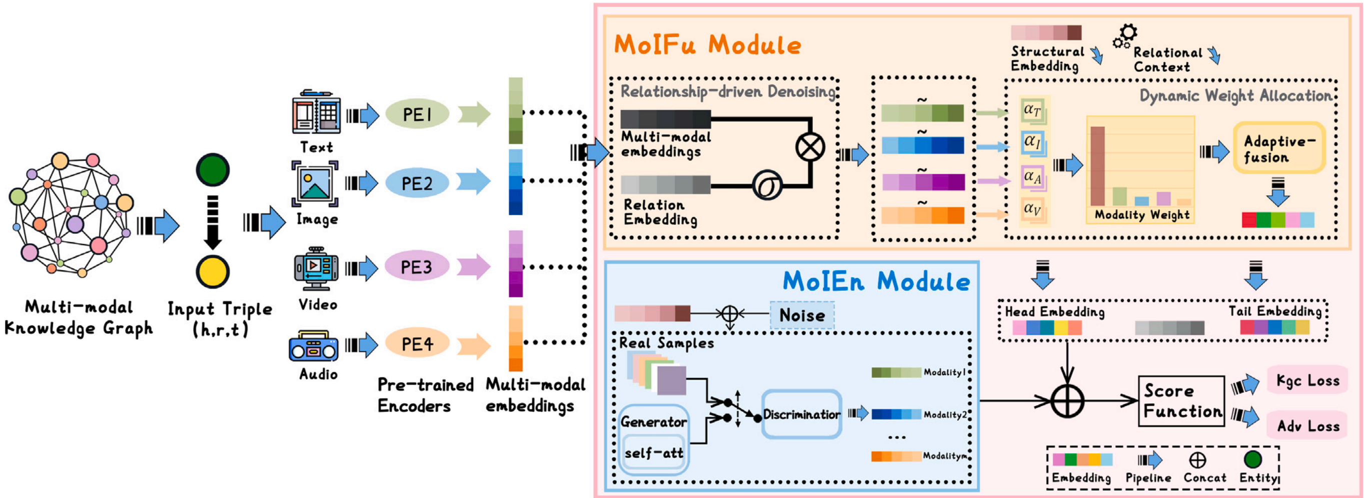


Fig. 2. The overall framework of the proposed AFME. AFME consists of two main modules: the Modal Information Fusion (MoIFu) module and the Modal Information Enhancement (MoIEn) module. MoIFu aims to dynamically fuse the semantic representations of different modalities through relationship-driven denoising and weight allocation. MoIEn focuses on enhancing modal information via adversarial generation mechanisms and refined feature optimization.

deep-layer features of each modality while preserving the uniqueness of modality-specific features. Finally, all modal features are unified into the same dimensional representation space through a two-layer projection, generating multi-modal embeddings.

4.2. *Modality information fusion*

The Modality Information Fusion (MoIFu) module adopts a relation-based denoising and weighted fusion approach. The relation-driven denoising mechanism helps suppress noise across modalities and balances the quality of different modalities. The dynamic weight allocation mechanism adaptively adjusts the modality weights based on feature quality and contextual relations, ensuring that high-quality modalities receive more attention during reasoning, thereby further improving the efficiency of multi-modal information fusion.

4.2.1. *Relationship-driven denoising*

Before multi-modal information fusion, a relationship-driven denoising module is designed in MoIFu to effectively reduce noise in modal features and enhance their representation quality. Unlike traditional denoising methods, this module focuses more on flexibly adjusting the noise suppression strength based on entity relations and inter-modal interactions, enabling it to better align with the link prediction task by reducing irrelevant noise while preserving useful information. This module uses the relationship embedding r as a guiding signal to dynamically control gating mechanisms, filtering out irrelevant or redundant features and retaining information closely related to the current relationship. Specifically, the relationship-driven denoising module generates a gating signal $gate_m$ for each modality. This signal is produced through the interaction between the modal feature h_m and the relationship embedding r, adjusting the feature passing ratio dynamically. The calculation of the gating signal is defined as follows:

$$\mathrm{gate}_\mathrm{m} = \sigma(\mathrm{W}_\mathrm{g} \cdot (\mathrm{h}_\mathrm{m} \odot \mathrm{r}) + \mathrm{b}_\mathrm{g}) \tag{4}$$

where σ represents the Sigmoid function, and W_g and b_g are the learnable weight matrix and bias term, respectively, used to perform linear transformation and generate control signals.⊙ denotes the element-wise multiplication operation, modeling the interaction between modal features and relationship embeddings. Subsequently, the gate signal is used to denoise the modal features, resulting in the refined modal features $\tilde{h}_\mathrm{m}$:

$$\tilde{h}_\mathrm{m} = \mathrm{gate}_\mathrm{m} \odot h_\mathrm{m} \tag{5}$$

The sanitized modal feature $\tilde{h}_\mathrm{m}$ retains information closely related to the current relationship context r while suppressing irrelevant or redundant noisy features, thereby enhancing the representational capability of the modal features. Through the relationship-driven denoising process, each modal feature is dynamically adjusted to be more refined and specific, providing high-quality input support for the subsequent dynamic weight allocation and modal information fusion. This is also an aspect that has not been thoroughly explored in existing research on knowledge graph link prediction.

4.2.2. *Dynamic weight allocation*

To address the imbalance in quantity and quality of modal information, AFME introduces an adaptive dynamic weight allocation mechanism in the Modal Information Fusion (MoIFu) module. Specifically, we design a dual weight allocation method based on confidence and contextual relationships. It can dynamically learn and adjust the fusion weights of each modality based on the internal quality indicator (confidence) of the modality features and their relevance to the current contextual relation. The input features of each modality can vary in quality (e.g., the image modality might suffer from low resolution or high noise, while the text modality might have incomplete semantics). To handle this, we calculate a “confidence” metric for each modality, determined by the L2 norm of the modal features, to dynamically adjust the modal weights. Modal features with higher quality will be given greater weight during the fusion process. Additionally, a relationship smoothing factor is used to further regulate the distribution of modal weights based on the current relationship. This ensures that the model focuses more on the relationship of the current triple, giving greater weight to the modal features that are more relevant to the relationship. This mechanism ensures that high-quality modality features receive more attention during the fusion process, while dynamically adapting to the feature requirements of different contextual relations, thereby improving the effectiveness of multi-modal information fusion and its context sensitivity. Compared with existing static weighted fusion methods, AFME’s dynamic weight allocation mechanism places greater emphasis on the relative importance of each modality under different conditions and adjusts accordingly based on the specific relational characteristics of the link prediction task, ultimately enhancing the model’s performance in multi-modal reasoning tasks.

For the triple(h,r,t), the weight of the head entity h in modality m is represented as:

$$\omega_\mathrm{m}(h, \mathrm{r}) = \frac{\exp\left(\alpha_\mathrm{m} \cdot \mathrm{U} \odot \tanh\left(\widetilde{h}_\mathrm{m}\right)/\sigma(\tau_\mathrm{r})\right)}{\sum_{\mathrm{n} \in \mathrm{M} \cup \{\mathrm{S}\}} \exp\left(\alpha_\mathrm{n} \cdot \mathrm{U} \odot \tanh\left(\widetilde{h}_\mathrm{n}\right)/\sigma(\tau_\mathrm{r})\right)} \tag{6}$$

where U is the learnable vector, τ_r is the relationship context smoothing factor, σ represents the Sigmoid function, and α_m is the confidence of modality m. Through this method, modality weights are dynamically adapted across different relationships and entities, enabling the fusion to generate joint embeddings.

4.3. Modality information enhancement

Inspired by the idea of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Papernot et al., 2016), this paper proposes an optimization strategy based on adversarial mechanisms in the Modality Information Enhancement (MoIEn) module. Although GANs have been widely applied in single-modality tasks such as image or text generation and data augmentation, existing methods still face certain limitations when applying GANs to multi-modal knowledge graph completion. The main challenge lies in the insufficient consideration of interactions and mutual influences among different modalities, which restricts the effectiveness of the generated features in semantic representation and completion reasoning. To address this, we adapt the GAN structure by introducing an optimization strategy that better captures the complex interactions among modality features, with specific improvements made to the design of the generator and discriminator to better meet the demands of multi-modal tasks.

4.3.1. Generator module

Traditional GAN generators typically produce new samples by taking random noise or existing data as input. However, this approach has certain limitations in the context of multi-modal knowledge graph completion. Due to the lack of guidance from structural knowledge and semantic relations between entities, modality completion based solely on noise or shallow features often lacks global consistency and semantic accuracy, leading to generated features that are disconnected from the original knowledge graph structure. To address this, we introduce global structural guidance into the generator, where the guiding signal is derived from entity relations within the knowledge graph. This is particularly important because structural modality plays a critical role in multi-modal knowledge graph tasks by providing deep semantic associations that help the generator more accurately complete missing information.

Specifically, the generator first combines structural modality embeddings with random noise to generate initial modality features. To improve the quality and consistency of these features, the generator further applies a multi-layer self-attention mechanism for deep refinement. In this process, self-attention captures both global and local dependencies within each modality, ensuring strong internal consistency of the generated features and enhancing the complementarity across different modalities. Through this optimization, the generator is able to produce high-quality modality features with stronger semantic consistency, which proves especially advantageous when dealing with complex multi-modal data and missing information.

(1) Structure Modality-Guided Initial Feature Generation: The generator takes the structural modality embedding e_s and the random noise vector z as inputs, combined with the initial representation $\widetilde{h}_m$ of the sanitized feature modality, to generate the preliminary representation $h_m^{(0)}$ of the feature modality. The specific steps are as follows:

First, the structural modality embedding e_s is concatenated with the random noise z to create a global guiding signal:

$$\mathrm{e_{mod}} = \mathrm{e_s} \oplus \mathrm{z} \tag{7}$$

where $\oplus$ represents the concatenation operation of vectors. Then, this guiding signal interacts with the representation of each feature modality$\widetilde{h}_m$, generating the initial step of guided feature modality representation:

$$\mathrm{h_m^{(0)}} = \sigma(\mathrm{W_d}(\mathrm{e_{mod}} \odot \tilde{\mathrm{h}}_\mathrm{m}) + \mathrm{b_d}) \tag{8}$$

where W_d and b_d are the specific transformation matrix and bias term of the generator, $\odot$ represents element-wise multiplication, used to realize the guidance of the structural modality for specific elements of the feature modality, and σ is the nonlinear activation function.

(2) Self-Attention Mechanism for Optimizing Modal Features: To fully explore the internal features of each modality, the initially generated modal features $h_m^{(0)}$ are fed into a multi-layer self-attention mechanism. This mechanism performs deep optimization of the internal information of each feature modality, capturing both global and local dependencies of the modal features and enhancing their representational capacity. The optimized features are computed as follows:

$$\mathrm{h_m^{(1)}} = \mathrm{softmax}\left(\left(\mathrm{h_m^{(0)}W_Q}\right)\left(\mathrm{h_m^{(0)}W_K}\right)^\top / \sqrt{\mathrm{d}}\right)\left(\mathrm{h_m^{(0)}W_V}\right) \tag{9}$$

where W_Q, W_K and W_V are the learnable projection matrices of the self-attention mechanism, and d is the scaling factor for the feature dimension.

(3) Inter-modal Interaction for Feature Optimization: After completing the optimization of single-modal internal features, the generator further optimizes the features through an inter-modal interaction mechanism, modeling the complementary relationships between different modalities. The specific formula is as follows:

$$\mathrm{h'_m} = \sum_{\mathrm{n}\neq\mathrm{m}} \alpha_{\mathrm{mn}} \mathrm{h_n^{(1)}} \tag{10}$$

where α_{mn} represents the interaction attention weight between modality m and modality n, defined as:

$$\alpha_{\mathrm{mn}} = \mathrm{softmax}\left(\mathrm{h_m^{(1)}} \cdot \mathrm{h_n^{(1)}} / \sqrt{\mathrm{d}}\right) \tag{11}$$

$h_n^{(1)}$ represents the optimized representation of the nth feature modality.

4.3.2. Discriminator module

The discriminator is designed to evaluate the degree of alignment between the generated modality embedding $\mathrm{h'_m}$ and the real modality embedding $\mathrm{h_m^{real}}$, ensuring that the features produced by the generator remain consistent with real modality features in both distribution and semantics. In multi-modal tasks, the interactions and dependencies among different modalities are often complex, and simply comparing generated and real features may not be sufficient to capture such patterns. Therefore, the discriminator not only guides the generator through standard adversarial training but also performs fine-grained evaluation of the generated features using multiple fully connected layers and self-attention mechanisms, further enhancing the quality of the generated modality features. First, the discriminator takes the generated features $\mathrm{h'_m}$ and the real features $\mathrm{h_m^{real}}$ as input and applies a self-attention mechanism to globally optimize the embeddings, extracting key feature patterns within the modalities. The optimized features are then mapped into the discriminator result space through fully connected layers, which are used to evaluate the authenticity of the generated features. The specific formula is as follows:

$$\mathrm{D}\left(\mathrm{h'_m}, \mathrm{h_m^{real}}\right) = \sigma\left(\mathrm{W_p} \cdot \left[\mathrm{h_m'^{(1)}} \parallel \mathrm{h_m^{real(1)}}\right] + \mathrm{b_p}\right) \tag{12}$$

where $\parallel$ denotes the concatenation operation, and $\mathrm{W_p}$ and $\mathrm{b_p}$ represent the weight matrix and bias term of the discriminator, respectively.

4.4. Model optimization and training objective

The training objective of the model consists of two main loss functions: the prediction loss and the adversarial generation loss. These two losses are optimized for different task objectives. The prediction loss is used to optimize the triple prediction task, aiming to improve the model's ability to evaluate the plausibility of triples. It enhances model performance by minimizing the error between predicted and true triples, thereby ensuring the accuracy and completeness of the knowledge graph. The adversarial generation loss is used to optimize the

Table 1
Statistical information of experimental datasets.

Dataset	Text		Image		Video		Audio		#Entity	#Relation	#Train	#Valid	#Test
	Num	Dim	Num	Dim	Num	Dim	Num	Dim					
MKG-W	14 123	384	14 463	383	–	–	–	–	15 000	169	34 196	4276	4274
MKG-Y	12 305	384	14 244	383	–	–	–	–	15 000	28	21 310	2665	2663
TIVA	11 858	300	11 636	2048	10 269	2048	2441	128	11 858	16	20 071	2000	2000
KVC16K	14 822	768	14 822	768	14 822	768	14 822	768	16 015	4	180 190	22 523	22 525

generative adversarial network, particularly the interaction between the generator and the discriminator. Its goal is to ensure that the generated modality features are semantically and structurally consistent with real features, while enhancing the model's robustness in multi-modal data scenarios. These two loss functions complement each other and jointly improve the model's performance in multi-modal knowledge graph completion.

4.4.1. Prediction task loss

In the multi-modal information fusion module, the final generated joint modal embeddings are used for the triple prediction task. The model evaluates the plausibility of triples(h,r,t) through a scoring function, and the specific formula is as follows:

$$F(h,r,t) = -\|h_{joint} \oplus r - t_{joint}\|_2 \tag{13}$$

where h_{joint} and t_{joint} represent the joint embeddings of the head and tail entities, r represents the relation embedding, $\oplus$ denotes the rotation operation in the complex space, and $\|\cdot\|_2$ is the L2 norm. To further enhance the model's triple prediction capability, we designed a loss function based on negative sampling:

$$\begin{aligned}\mathcal{L}_{KGC} = &-\frac{1}{|T|}\sum_{(h,r,t)\in T}\log\sigma\left(\delta + F(h,r,t)\right)\\ &-\frac{1}{|T'|}\sum_{(h',r,t')\in T'}\log\sigma\left(-\delta - F\left(h',r,t'\right)\right)\end{aligned} \tag{14}$$

where T represents the set of real triples, T' represents the set of virtual triples generated through negative sampling, σ is the Sigmoid activation function, and δ is a fixed margin hyperparameter.

4.4.2. Adversarial loss

To further improve the stability of adversarial training between the generator and the discriminator, AFME adopts the loss design concept from WGAN. Specifically, the goal of the generator is to produce modality features that are as close as possible to the real distribution, making it difficult for the discriminator to distinguish between real and generated samples. The discriminator, on the other hand, aims to enhance its discrimination capability by maximizing the difference between real and generated modality features. This optimization process is formulated as a min–max loss function, as shown in the following equation, where the core objective is to approximate the target modality distribution through adversarial optimization, thereby achieving distribution alignment and modality completion.

$$\begin{aligned}\underset{G}{\min}\max L_{\mathrm{Adv}} = &\mathbb{E}_{\mathrm{h}_{\mathrm{m}}^{\mathrm{real}}\sim P_{\mathrm{real}}}[\log D(\mathrm{h}_{\mathrm{m}}^{\mathrm{real}})]\\ &+\mathbb{E}_{\mathrm{h}'_{\mathrm{m}}\sim P_G}\left[\log\left(1 - D\left(\mathrm{h}'_{\mathrm{m}}\right)\right)\right]\end{aligned} \tag{15}$$

However, in practical training, relying solely on traditional adversarial loss may lead to issues such as gradient vanishing or model instability. To address this, inspired by the optimization strategy of WGAN-GP, we introduce a gradient penalty term L_{gp} (Gulrajani et al., 2017). This term constrains the norm of the gradient of the discriminator's output with respect to its input features, enforcing the 1-Lipschitz continuity condition. As a result, it effectively mitigates gradient explosion or vanishing problems, and improves both the convergence speed and training stability of the model. Its definition is as follows:

$$L_{\mathrm{gp}} = \lambda\mathbb{E}_{\tilde{e}_m\sim P_{\tilde{e}_m}}\left[\left(\|\nabla_{\tilde{e}_m}D(\tilde{e}_m)\|_2 - 1\right)^2\right] \tag{16}$$

where $\tilde{e}_m$ represents the interpolated samples between the generated modal embedding $\mathrm{h}_{\mathrm{m}}^{'}$ and the real modal embedding $\mathrm{h}_{\mathrm{m}}^{\mathrm{real}}$ and λ is the balance coefficient for the gradient penalty term, used to adjust the strength of the gradient constraint.Compared with the previous approach of weight clipping in the discriminator, this method provides a more stable way to control the training dynamics, ensuring that it does not interfere with the primary optimization objectives of the generator and the discriminator.

4.4.3. Final objective

The final objective function of AFME integrates the negative sampling loss L_{kgc} for the knowledge graph completion task and the adversarial training loss L_{adv} for the generator-discriminator. The overall optimization objective is:

$$L = L_{\mathrm{kgc}} + \lambda_1 L_{\mathrm{adv}} + \lambda_2 L_{\mathrm{gp}} \tag{17}$$

Here, λ_1 and λ_2 are weighting coefficients, and L_{gp} is the gradient penalty term, used to stabilize the training of the adversarial network. Through the above optimization objective, AFME achieves effective reasoning and completion within the complexity, diversity, and imbalance of multi-modal information.

5. Experiments

In this section, we conduct a comprehensive evaluation of the performance of the AFME framework through a series of experiments. We first introduce the experimental setup and then discuss the effectiveness and advantages of the framework in the multi-modal knowledge graph completion task. Specifically, we aim to address the following key research questions (RQs):

- RQ1: Can AFME outperform existing baseline methods and achieve significant performance improvements in the multi-modal knowledge graph completion task?
- RQ2: How much does each module of AFME contribute to the final performance?
- RQ3: Does the MoIEn strategy demonstrate good generalizability? How does the performance of different models improve after incorporating this module?
- RQ4: How does AFME's performance vary with changes in key hyperparameters? Under different hyperparameter settings, which values maximize the model's completion performance?

5.1. Experimental setup

5.1.1. Datasets

To better explore the MMKGC task in more complex and diverse environments, we conducted extensive experiments on four publicly available datasets. Our benchmarks include the following datasets:

(1) MKG-W (Xu et al., 2022): An MMKG derived from Wikidata (Vrandečić & Krötzsch, 2014), containing text and image modalities (Xu et al., 2022).

(2) MKG-Y (Xu et al., 2022): An MMKG derived from YAGO (Suchanek, Kasneci, & Weikum, 2007), containing text and image modalities (Xu et al., 2022).

Table 2
Performance comparison of different MMKGC models.

Method	MKG-W		MKG-Y		TIVA			KVC16K		
	MRR	Hits@1	MRR	Hits@1	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
TransE	29.19	21.06	30.73	23.45	83.85	83.20	84.15	8.54	0.64	23.42
DistMult	20.99	15.93	25.04	19.33	82.27	81.15	84.22	6.37	3.03	12.61
ComplEx	24.93	19.09	28.71	22.26	80.67	77.67	86.10	12.85	7.48	23.18
RotatE	33.67	26.80	34.95	29.10	84.59	83.47	86.95	14.33	8.25	26.17
IKRL	32.36	26.11	33.22	30.37	67.71	63.72	75.67	11.11	5.42	22.39
TBKGC	31.48	25.31	33.99	30.47	81.57	78.75	86.05	5.39	0.35	15.52
TransAE	30.00	21.23	28.10	25.31	79.57	74.95	88.07	10.81	5.31	21.89
MMKRL	30.10	22.16	36.81	31.66	85.03	81.92	90.10	8.78	3.89	18.34
RSME	29.23	23.36	34.44	31.78	40.01	30.55	51.35	12.31	7.14	22.05
VBKGC	30.61	24.91	37.04	33.76	74.07	66.87	85.85	14.66	8.28	27.04
OTKGE	34.36	28.85	35.51	31.97	35.28	30.45	41.98	8.77	5.01	15.55
IMF	34.50	28.77	35.79	32.95	55.46	41.87	77.57	12.01	7.42	21.01
QEB	32.38	25.47	34.37	29.49	74.25	66.10	88.35	12.06	5.57	25.01
VISTA	32.91	26.12	30.45	24.87	76.07	70.67	86.60	11.89	6.97	21.27
KBGAN	29.47	22.21	29.71	22.81	85.44	82.45	90.10	13.72	7.54	25.88
MANS	30.88	24.89	29.03	25.25	85.70	82.70	90.62	10.42	5.21	20.45
MMRNS	35.03	28.59	35.93	30.53	83.12	83.05	83.25	13.31	7.51	24.68
NATIVE	36.48	29.35	38.81	34.67	91.61	90.80	92.43	15.41	9.12	28.00
AFME	**37.09**	**30.33**	**39.41**	**35.45**	**92.44**	**91.64**	**93.70**	**15.83**	**9.32**	**28.64**
Improvement	+1.67%	+3.33%	+1.55%	+2.25%	+0.91%	+0.93%	+1.37%	+2.73%	+2.19%	+2.29%

(3) KVC16K (Pan et al., 2022): This dataset is adapted from the video concept encyclopedia KuaiPedia (Pan et al., 2022), containing four modalities: text, image, video, and audio.

(4) TIVA (Wang et al., 2023): This is an MMKG that includes four modalities: text, image, video, and audio.

Detailed statistical information about the datasets is shown in Table 1.

5.1.2. Evaluation metrics

Following prior work, we use ranking-based metrics (Sun et al., 2019) to evaluate the triple prediction task (Bordes et al., 2013), including Mean Reciprocal Rank (MRR) and Hit Rates (Hits@1 and Hits@10). Most existing studies also primarily rely on ranking-based metrics as the standard for performance evaluation, which is widely accepted in current multi-modal knowledge graph completion tasks. To ensure fairness in comparison, we adopt a filtered setting (Bordes et al., 2013), removing candidate triples that already exist in the training set. The final results are computed by averaging the performance of head entity prediction and tail entity prediction.

5.1.3. Baseline comparison

To evaluate the effectiveness of our proposed model, we perform a comprehensive comparison with representative models in the KGC domain. The baselines are broadly divided into three categories:

(1) Traditional Single-modal KGC Methods: These methods design simple scoring functions and learn structural embeddings for a given KG without using any multi-modal information.

- TransE (Bordes et al., 2013): Models entities and relationships in a knowledge graph through vector translation, aiming to make the vector sum of the head entity and relationship close to the tail entity's vector.
- DistMult (Yang et al., 2014): Uses a bilinear scoring function to capture the symmetry and semantic information of relationships, performing well under single-modal conditions.
- ComplEx (Trouillon et al., 2016): Extends the DistMult model by modeling entities and relationships in the complex vector space, allowing it to capture more complex relationship patterns.
- RotatE (Sun et al., 2019): Models relationships using rotational operations in the complex vector space, enabling it to represent more intricate relationship types.

(2) MMKGC Methods: These methods consider both multi-modal information and triple structural information.

- IKRL (Xie et al., 2016): Incorporates multi-modal information (e.g., images and text) and completes knowledge graphs using an attention mechanism-based approach.
- TBKGC (Mousselly-Sergieh et al., 2018): Combines multi-modal representations with knowledge graph embeddings, focusing on addressing the inconsistency of multi-modal information.
- TransAE (Wang et al., 2019): Introduces autoencoders into knowledge graph embeddings to capture latent features from multi-modal information.
- MMKRL (Lu et al., 2022): Proposes a joint learning framework that simultaneously optimizes multi-modal embeddings and knowledge reasoning tasks.
- RSME (Wang, Wang et al., 2021): Integrates multi-modal features using sparse coding techniques, improving the representation capabilities of multi-modal information.
- VBKGC (Zhang & Zhang, 2022): Combines visual features as complementary information with knowledge graph embeddings for fine-grained entity modeling.
- OTKGE (Cao et al., 2022): Adopts an optimal transport mechanism for weighted fusion of modal features, enabling dynamic adjustments of different modal features.
- IMF (Li et al., 2023): Introduces a modal fusion mechanism to comprehensively consider the complementarity and correlations between modal features.
- QEB (Wang et al., 2023): Designs a query embedding-based model that integrates modal information to achieve more precise reasoning.
- VISTA (Lee et al., 2023): Leverages visual and textual features to support knowledge graph tasks, enhancing completion performance through cross-modal learning.

(3) Negative Sampling Methods: These methods leverage multi-modal information to generate high-quality negative samples, improving the training process.

- KBGAN (Wang, Shen et al., 2021): Combines generative adversarial networks and reinforcement learning to optimize the negative sample generation strategy.
- MANS (Zhang et al., 2023): Designs a multi-modal negative sampling method to generate higher-quality negative samples and improve model performance.
- MMRNS (Xu et al., 2022): Enhances negative samples by incorporating multi-modal correlations, improving training efficiency.

- NATIVE (Zhang et al., 2024): Develops an adaptive modal fusion method and uses adversarial networks to increase the contribution of negative samples to model optimization.

5.1.4. Implementation details

To ensure the stability of the model training process and the reproducibility of results, a series of standard optimization strategies and parameter settings were adopted during implementation and training. All experiments were conducted on an Ubuntu 20.04 system. Computation was performed on an NVIDIA RTX 4090 GPU. During training, the Adam optimizer was used for parameter updates, as it provides stable convergence when dealing with sparse gradients and non-stationary objectives. In the adversarial training stage, both the generator and the discriminator used Adam optimizers, and the learning rate was tuned from the set {1e−3, 1e−4, 1e−5}. The batch size was set to 1024, and the training was capped at 1000 epochs. An early stopping mechanism was employed to prevent overfitting. For triple modeling, the margin value between positive and negative samples was tuned within the range [2, 6], and was finally set to 4. The number of negative samples per positive instance was selected from {32, 64, 128}, with 64 providing a good trade-off between training efficiency and performance. In the modality information enhancement module, the dimension of the random noise input to the generator was set to 64, which is used to complete missing or incomplete modality information. For the adversarial loss function, two balancing coefficients, and, were tuned from the set {1e−2, 1e−3, 1e−4, 1e−5}.

5.2. Performance comparison (RQ1)

Table 2 presents the performance comparison of our AFME framework with other baseline methods across four datasets. In Table 2, AFME and the best-performing baseline methods are highlighted in bold and underlined, respectively.

(1) The experimental results show that AFME outperforms all existing baseline methods across all datasets, achieving state-of-the-art performance. Specifically: Compared to the most advanced baseline method, NATIVE, AFME achieves MRR improvements of 1.67%, 1.55%, 0.88%, and 2.73% on the MKG-W, MKG-Y, TIVA, and KVC16K datasets, respectively. For the Hits@1 metric, AFME achieves increases of 3.33%, 2.25%, 0.86%, and 2.19% on the same datasets. Additionally, for the Hits@10 metric, AFME achieves improvements of 1.37% and 2.29% on the TIVA and KVC16K datasets, respectively. These results strongly validate AFME's significant advantages in addressing the challenges of modality complexity, diversity, and quality imbalance in multi-modal knowledge graph completion tasks.

(2) Further analysis reveals that even when using a limited amount of modality information typically considered by mainstream methods (such as structural, textual, and image modalities), AFME still outperforms baseline methods (e.g., IKRL and VBKGC). This demonstrates that AFME is capable of effectively extracting and optimizing modality features under limited modality information. Moreover, when compared with multi-modal information and adversarial-based methods (e.g., MMRNS and NATIVE), AFME consistently achieves better performance. This further validates the superiority of the adversarial generation module we designed in enhancing modality information and filling in missing information. It effectively generates reasonable missing modality features and significantly improves the overall reasoning capability.

5.3. Ablation study (RQ2)

To validate the effectiveness of each module in AFME, we conducted a series of ablation experiments to evaluate the contribution of each module to the overall performance. Specifically, we constructed variant models by removing key components, such as the relationship-driven noise reduction and dynamic weight allocation in the MoIFu module, as well as the information enhancement optimization in the MoIEn module. Two sets of experiments were conducted on the TIVA and MKG-Y datasets. The detailed results are shown in Fig. 3 and Table 3.

Table 3
Ablation study of key components in AFME.

Datasets	Variants	MRR	Hits@1	Hits@10
MKG-Y	AFME	39.41	35.45	–
	AFME$_{w/o\text{-}MoIFu}$	38.87	34.70	–
	AFME$_{w/o\text{-}MoIEn}$	37.33	33.35	–
TIVA	AFME	92.42	91.58	93.7
	AFME$_{w/o\text{-}MoIFu}$	91.99	90.9	93.08
	AFME$_{w/o\text{-}MoIEn}$	90.75	89.55	92.18

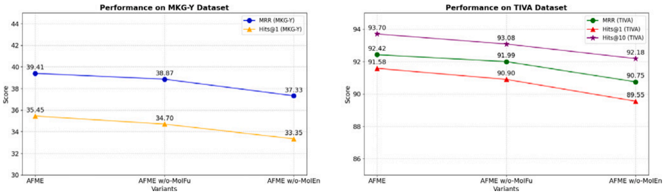


Fig. 3. Performance comparison of different AFME variants.

We set up the following variants:

(1) w/o-MoIFu: The relationship-driven noise reduction and weight allocation module in the modality fusion part was removed. Instead, the extracted feature vectors from multi-modalities were directly concatenated for fusion.

(2) w/o-MoIEn: The key intra-modal self-attention mechanism in the modality information enhancement part was removed. The generator directly generated pseudo-features, lacking the guidance of structural modality and the complex interaction process.

From the analysis of the ablation experiment data, the following conclusions can be drawn:

(1) The MoIFu module significantly reduces noise interference in modality features through a relation-driven denoising mechanism and a dynamic weight allocation strategy, while effectively alleviating the imbalance in modality information quality. Experimental results show that simple feature concatenation methods fail to fully exploit the complementarity between modalities, which leads to a notable decline in the efficiency of multi-modal feature utilization. In contrast, MoIFu achieves a better balance of contributions from different modalities during information fusion, thereby improving overall performance.

(2) The MoIEn module performs deep optimization of intra-modal features using a self-attention mechanism and achieves deep intermodal interaction and completion by guiding feature modalities with structural modality. Experimental results show that a generator without self-attention and structural guidance fails to effectively optimize features, resulting in a significant decline in the authenticity and consistency of the generated features. Through an effective optimization process, the MoIEn module improves the quality of generated features, making them more aligned with the semantic structure of real modality features.

(3) Synergy of MoIFu and MoIEn Modules in AFME Framework: The collaborative interaction of the MoIFu and MoIEn modules plays a crucial role in improving the performance of multi-modal knowledge graph completion. The MoIFu module addresses noise interference and information quality imbalance during the modality feature fusion process, laying the foundation for efficient modality integration. Meanwhile, the MoIEn module further enhances deep interaction between modalities and optimizes the process of generating features, leading to significant overall performance improvement.

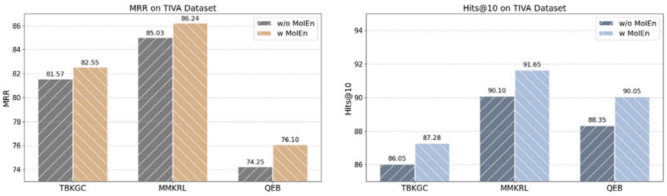


Fig. 4. Generalization experiment of the MoIEn module on three different MMKGC models. We report the results for MRR and Hits@1 metrics on the TIVA dataset.

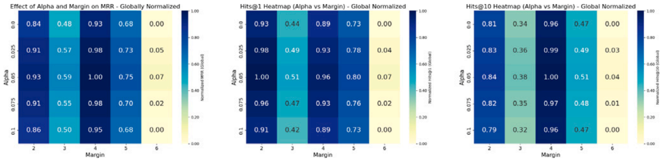


Fig. 5. Impact of relationship-driven noise reduction intensity (α) and positive–negative sample margin (Margin) on overall performance.

5.4. Generalization experiment (RQ3)

To verify whether the proposed MoIEn module is a generalizable framework, we designed a set of experiments by integrating this module into other representative baseline models (such as TBKGC, MMKRL, and QEB) and testing its performance improvement on the multi-modal dataset TIVA. The results are shown in Fig. 4.

The experimental results show that integrating the MoIEn module into other baseline models leads to performance improvements across the board. The extent of the improvement is closely related to the complexity of the baseline model itself. This further validates the capability of the MoIEn module to optimize intra-modal features through self-attention mechanisms and achieve deep inter-modal interactions using generative adversarial networks, demonstrating its strong adaptability. Moreover, the results also highlight the module's cross-model transferability and stability, as it consistently enhances performance across different models without requiring significant modifications. This reinforces its potential as a generalizable and robust enhancement framework for multi-modal knowledge graph completion tasks.

5.5. Sensitivity analysis (RQ4)

In this section, we investigate the sensitivity of the AFME model to different hyperparameters on the TIVA dataset and further analyze the rationale behind the selected hyperparameter values based on experimental results, explaining their impact on model performance.

(1) Controlling Relationship-Driven Noise Reduction Intensity α: In the modality fusion module, the intensity of relationship-driven noise reduction is controlled by the hyperparameter α. This parameter balances the proportion of raw modality information and purified features to achieve adaptive noise suppression. Specifically, when α approaches 0, the model relies more on raw modality features and ignores the noise reduction process; when α approaches 1, the model tends to rely entirely on the purified modality features. To analyze the impact of α on model performance, we set it to five different values and conducted experiments on the TIVA dataset. The results, as shown in Fig. 5. It show that a moderate denoising strength can effectively balance noise filtering and information retention, thereby improving the quality of feature representations. When α is too small (e.g., 0) or too large (e.g., 0.1), the model performance degrades. A smaller α leads to insufficient noise removal, while a larger α may overly simplify the original modality information and result in the loss of important features. Therefore, selecting an appropriate denoising strength is crucial for improving model performance.

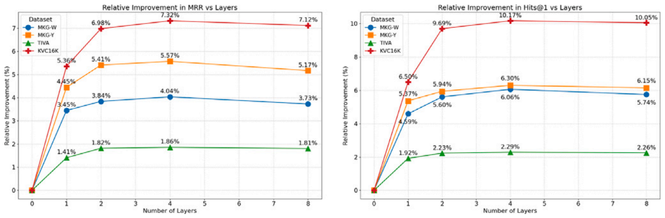


Fig. 6. Impact of the number of self-attention layers on different datasets.

(2) Controlling the positive–negative sample distance with Margin:This parameter is used to regulate the minimum separation distance between positive and negative samples, which plays a crucial role in the learning effectiveness of the triple scoring function. A smaller margin may lead to insufficient discriminative ability, causing the model to confuse positive and negative samples. In contrast, a larger margin can increase the score gap but may result in overfitting or unstable training. To analyze the impact of this parameter on model performance, we conducted multiple experiments on the TIVA dataset. As shown in Fig. 5, the model achieved the best performance when the margin was set to 4, indicating that a moderate separation distance helps strike a good balance between discrimination capability and training stability, thereby improving the overall performance of the model.

(3) Number of self-attention layers: In the modality information enhancement module, the number of self-attention layers determines the model's ability to perform deep optimization of modality features. To investigate the impact of different layer numbers on model performance, we set the number of layers to 0, 1, 2, 4, and 8, and conducted experiments on our benchmark dataset. The results, as shown in Fig. 6, indicate that the optimal number of self-attention layers is 4. At this level, the model effectively captures the complex interactions and global dependencies of modality features at an appropriate depth. When the number of layers is too small, the model fails to effectively capture the complex relationships among multi-modal features. On the other hand, using too many layers may lead to excessive computational complexity and potential overfitting in some cases. Therefore, we choose 4 layers as the optimal number for the self-attention mechanism to achieve a good balance between efficiency and performance.

Through the above experiments and analysis, we verified that the proposed method maintains stable performance under most hyperparameter settings, with significant performance degradation occurring only under extreme configurations. We demonstrated the impact of each selected hyperparameter on model performance and explained the rationale behind the chosen optimal values. The results not only confirm the effectiveness of the hyperparameter settings but also provide useful guidance for future parameter tuning in similar multi-modal learning tasks.

6. Conclusion

This paper proposes a novel multi-modal knowledge graph completion framework, AFME, to address challenges such as the complexity and diversity of multi-modal information, imbalance among modalities, and insufficient inter-modal interaction. The AFME framework achieves comprehensive optimization of multi-modal knowledge graph representation and reasoning through the synergistic interaction of its modality information fusion module (MoIFu) and modality information enhancement module (MoIEn). In MoIFu, relationship-driven noise reduction and dynamic weight allocation are implemented, enabling effective fusion of multi-modal features. In MoIEn, generative adversarial networks are used to complete missing modality features, while multi-layer self-attention mechanisms and discriminators enhance the authenticity and consistency of modality features. Experimental results demonstrate that the proposed framework exhibits excellent robustness in handling complex scenarios such as modality missing and noise interference.

Despite AFME's strong performance in multi-modal knowledge graph completion tasks, it lacks in-depth analysis of generalization across different relation types and reasoning over rare entities. The standard datasets used provide a solid foundation but mainly focus on common entities and relations, with limited coverage of complex relational structures and low-frequency entities, restricting validation in more realistic scenarios. To address these issues, Future research can further explore evaluating and optimizing AFME on more challenging datasets, such as medical VQA and social network datasets, to improve its generalization and robustness. Additionally, investigating more efficient modality representation learning methods to reduce computational costs is crucial for practical applications. We believe AFME's design offers new insights and broad applicability in knowledge reasoning and other complex real-world tasks.

CRediT authorship contribution statement

Zenglong Wang: Writing – review & editing. **Xuan Liu:** Writing – review & editing. **Zheng Liu:** Writing – review & editing. **Yu Weng:** Writing – review & editing. **Chaomurilige:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities, China under Grant 2024QNYL07, in part by the National Key Research and Development Program of Hainan Province, China under Grant ZDYF2024(LALH)005, and in part by the Beijing Municipal Science and Technology Commission, China under Grant Z231100001723002. Zenglong Wang, Xuan Liu, Zheng Liu, Yu Weng and Chaomurilige are with the Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance, Minzu University of china, Beijing, China.

References

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, *26*.

Cao, Z., Xu, Q., Yang, Z., He, Y., Cao, X., & Huang, Q. (2022). Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems*, *35*, 39090–39102.

Chen, Z., Fang, Y., Zhang, Y., Guo, L., Chen, J., Chen, H., et al. (2024). The power of noise: Toward a unified multi-modal knowledge graph representation framework. arXiv preprint arXiv:2403.06832.

Chen, Z., Guo, L., Fang, Y., Zhang, Y., Chen, J., Pan, J. Z., et al. (2023). Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *International semantic web conference* (pp. 121–139). Springer.

Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, *141*, Article 112948.

Chen, L., & Li, K. (2025). Multi-modal graph aggregation transformer for image captioning. *Neural Networks*, *181*, Article 106813.

Chen, Z., Zhang, Y., Fang, Y., Geng, Y., Guo, L., Chen, X., et al. (2024). Knowledge graphs meet multi-modal learning: A comprehensive survey. arXiv preprint arXiv:2402.05391.

Croce, D., Castellucci, G., & Basili, R. (2020). GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2114–2119).

Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. *vol. 32*, In *Proceedings of the AAAI conference on artificial intelligence*. 1.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, *30*.

Jenatton, R., Roux, N., Bordes, A., & Obozinski, G. R. (2012). A latent factor model for highly multi-relational data. *Advances in Neural Information Processing Systems*, *25*.

Karen, S. (2014). Very deep convolutional networks for large-scale image recognition. ArXiv Preprint arXiv:1409.1556.

Karras, T. (2019). A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *vol. 1*, In *Proceedings of naacL-HLT* (p. 2). Minneapolis, Minnesota.

Lee, J., Chung, C., Lee, H., Jo, S., & Whang, J. (2023). VISTA: Visual-textual knowledge graph representation learning. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 7314–7328).

Li, X., Zhao, X., Xu, J., Zhang, Y., & Xing, C. (2023). IMF: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM web conference 2023* (pp. 2572–2580).

Liang, K., Liu, Y., Zhou, S., Tu, W., Wen, Y., Yang, X., et al. (2023). Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, *36*(1), 226–238.

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., et al. (2023). Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval* (pp. 1559–1568).

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., et al. (2024). A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liang, W., Meo, P. D., Tang, Y., & Zhu, J. (2024). A survey of multi-modal knowledge graphs: Technologies and trends. *ACM Computing Surveys*, *56*(11), 1–41.

Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., & Rosenblum, D. S. (2019). MMKG: Multi-modal knowledge graphs. In *The semantic web: 16th international conference, ESWC 2019, portorož, Slovenia, June 2–6, 2019, proceedings 16* (pp. 459–474). Springer.

Lu, X., Wang, L., Jiang, Z., He, S., & Liu, S. (2022). MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1–18.

Mescheder, L., Nowozin, S., & Geiger, A. (2017). The numerics of gans. *Advances in Neural Information Processing Systems*, *30*.

Mousselly-Sergieh, H., Botschen, T., Gurevych, I., & Roth, S. (2018). A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the seventh joint conference on lexical and computational semantics* (pp. 225–234).

Nayyeri, M., Cil, G. M., Vahdati, S., Osborne, F., Rahman, M., Angioni, S., et al. (2021). Trans4E: Link prediction on scholarly knowledge graphs. *Neurocomputing*, *461*, 530–542.

Pan, H., Zhai, Z., Zhang, Y., Fu, R., Liu, M., Song, Y., et al. (2022). Kuaipedia: A large-scale multi-modal short-video encyclopedia. arXiv preprint arXiv:2211.00732.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (euroS&p)* (pp. 372–387). IEEE.

Pezeshkpour, P., Chen, L., & Singh, S. (2018). Embedding multimodal relational data for knowledge base completion. arXiv preprint arXiv:1809.01341.

Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(2), 1–49.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* (pp. 697–706).

Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In *International conference on machine learning* (pp. 2071–2080). PMLR.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85.

Wang, Z., Li, L., Li, Q., & Zeng, D. (2019). Multimodal data enhanced representation learning for knowledge graphs. In *2019 international joint conference on neural networks* (pp. 1–8). IEEE.

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, *29*(12), 2724–2743.

Wang, X., Meng, B., Chen, H., Meng, Y., Lv, K., & Zhu, W. (2023). TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 2391–2399).

Wang, B., Shen, T., Long, G., Zhou, T., Wang, Y., & Chang, Y. (2021). Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the web conference 2021* (pp. 1737–1748).

Wang, M., Wang, S., Yang, H., Zhang, Z., Chen, X., & Qi, G. (2021). Is visual context really helpful for knowledge graph? A representation learning perspective. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 2735–2743).

Wang, X., Xu, Y., He, X., Cao, Y., Wang, M., & Chua, T.-S. (2020). Reinforced negative sampling over knowledge graph for recommendation. In *Proceedings of the web conference 2020* (pp. 99–109).

Wang, J., Yu, L., Zhang, W., Gong, Y., Xu, Y., Wang, B., et al. (2017). Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 515–524).

Wei, W., Huang, C., Xia, L., & Zhang, C. (2023). Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM web conference 2023* (pp. 790–800).

Xie, R., Liu, Z., Luan, H., & Sun, M. (2016). Image-embodied knowledge representation learning. arXiv preprint arXiv:1609.07028.

Xu, D., Xu, T., Wu, S., Zhou, J., & Chen, E. (2022). Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 3857–3866).

Yang, B., Yih, W.-t., He, X., Gao, J., & Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.

Yao, L., Mao, C., & Luo, Y. (2019). KG-BERT: BERT for knowledge graph completion. arXiv preprint arXiv:1909.03193.

Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. *vol. 31*, In *Proceedings of the AAAI conference on artificial intelligence*. 1.

Zhang, Y., Chen, Z., Guo, L., Xu, Y., Hu, B., Liu, Z., et al. (2024). Native: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval* (pp. 91–101).

Zhang, Y., Chen, M., & Zhang, W. (2023). Modality-aware negative sampling for multi-modal knowledge graph embedding. In *2023 international joint conference on neural networks* (pp. 1–8). IEEE.

Zhang, H., Gong, Y., Shen, Y., Lv, J., Duan, N., & Chen, W. (2021). Adversarial retriever-ranker for dense text retrieval. arXiv preprint arXiv:2110.03611.

Zhang, Q., Wang, R., Yang, J., & Xue, L. (2022). Structural context-based knowledge graph embedding for link prediction. *Neurocomputing*, *470*, 109–120.

Zhang, Y., & Zhang, W. (2022). Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. arXiv preprint arXiv:2209.07084.

Zhu, Z., Zhang, Z., Xhonneux, L.-P., & Tang, J. (2021). Neural bellman-ford networks: A general graph neural network framework for link prediction. *Advances in Neural Information Processing Systems*, *34*, 29476–29490.