Collaboration of Fusion and Independence: Hypercomplex-driven Robust Multi-Modal Knowledge Graph Completion

Zhiqiang Liu[1,3], **Yichi Zhang**[2,3], **Mengshu Sun**[4], **Lei Liang**[4], **Wen Zhang**[1,3]*
[1]School of Software Technology, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]ZJU-Ant Group Joint Lab of Knowledge Graph
[4]Ant Group
{zhiqiangliu,zhang.wen}@zju.edu.cn

Abstract

Multi-modal knowledge graph completion (MMKGC) aims to discover missing facts in multi-modal knowledge graphs (MMKGs) by leveraging both structural relationships and diverse modality information of entities. Existing MMKGC methods follow two multi-modal paradigms: fusion-based and ensemble-based. Fusion-based methods employ fixed fusion strategies, which inevitably leads to the loss of modality-specific information and a lack of flexibility to adapt to varying modality relevance across contexts. In contrast, ensemble-based methods retain modality independence through dedicated sub-models but struggle to capture the nuanced, context-dependent semantic interplay between modalities. To overcome these dual limitations, we propose a novel MMKGC method **M-Hyper**, which achieves the coexistence and collaboration of fused and independent modality representations. Our method integrates the strengths of both paradigms, enabling effective cross-modal interactions while maintaining modality-specific information. Inspired by "quaternion" algebra, we utilize its four orthogonal bases to represent multiple independent modalities and employ the Hamilton product to efficiently model pair-wise interactions among them. Specifically, we introduce a Fine-grained Entity Representation Factorization (FERF) module and a Robust Relation-aware Modality Fusion (R2MF) module to obtain robust representations for three independent modalities and one fused modality. The resulting four modality representations are then mapped to the four orthogonal bases of a biquaternion for comprehensive modality interaction. Extensive experiments indicate its state-of-the-art performance with better robustness. Our dataset and code are available at https://github.com/zjukg/M-Hyper.

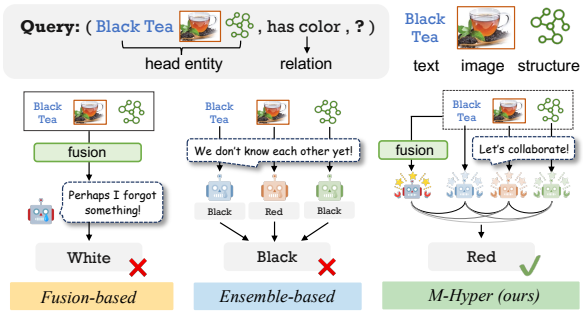

Figure 1: A simple example illustrates the difference between M-Hyper and existing paradigms.

1 Introduction

Multi-modal Knowledge Graphs (MMKGs) (Liu et al., 2019) expand traditional knowledge graphs by incorporating additional multi-modal information, making them more powerful tools (Chen et al., 2024) for knowledge representation. This makes MMKGs valuable for various applications, including recommendation systems (Wang et al., 2019a) and natural language processing (Chen et al., 2023b; Liu et al., 2025). However, like traditional uni-modal knowledge graphs (Liu et al., 2024), MMKGs also suffer from incomplete information (Xie et al., 2017); this limitation has been ameliorated through Multi-Modal Knowledge Graph Completion (MMKGC) methods.

As shown in Figure 1, existing MMKGC approaches fall into two paradigms: fusion-based and ensemble-based. Fusion-based methods (Zhang et al., 2025a) achieve cross-modality interaction via explicit fusion modules or dedicated cross-modality loss functions. Yet, their reliance on fixed fusion strategies often leads to suboptimal representation: crucial unique modality cues can be lost during fusion, and the model struggles to flexibly adapt to varying modality salience and synergies required in distinct reasoning contexts. Conversely, ensemble-based methods (Li et al., 2023)

*Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27968–27983

preserve modality-specific characteristics by employing independent sub-models, but inevitably fail to capture subtle inter-modal dependencies and interactions that are critical for complex reasoning scenarios. This highlights a fundamental challenge: the modality requirements in MMKGs exhibit dynamic, context-dependent, and task-specific contributions, making rigid adherence to either independent or fully fused paradigms a significant limitation to the expressive power and adaptability of MMKGC models. Hence, we propose the following research question: is it possible to develop a method that **combines the strengths of both paradigms, adapting to both fused and independent modality requirements while dynamically enabling comprehensive cross-modal interactions?**

To address these limitations, we introduce **M-Hyper**, the first method to model MMKGs in a **hyper**complex space. Inspired by quaternion algebra, where the four orthogonal basis elements preserve linear independence, M-Hyper explicitly separates distinct modality representations to retain original modal information and leverages the Hamilton product to facilitate comprehensive pairwise interactions among modalities. To enhance the robustness of modality representations, we design two novel modules: Fine-grained Entity Representation Factorization (FERF), which yields robust representations for three independent modalities, and Robust Relation-aware Modality Fusion (R2MF), which produces one robust fused modality representation. These four representations are mapped to the four orthogonal bases of a biquaternion, and a biquaternion-based scoring function is used to fully capture cross-modal semantic information. Experimental results show that our M-Hyper achieves state-of-the-art performance on three MMKGC datasets and exhibits high robustness and computational efficiency. Our contributions can be summarized as follows:

- We highlight the limitations of existing MMKGC paradigms and propose a novel biquaternion-based representation approach that simultaneously preserves both individual and fused modalities.
- We propose M-Hyper, the first MMKGC method operating in a hypercomplex (biquaternion) space, enabling robust coexistence and collaboration of fused and independent modality representations.
- Extensive empirical evaluation on three MMKGC benchmarks demonstrates that M-Hyper outperforms 18 existing baseline methods, exhibiting superior robustness and computational efficiency.

2 Related Works

2.1 Hypercomplex-based KG Embedding

Knowledge graph embedding (KGE) aims to project entities and relations into continuous vector spaces to capture complex relational patterns. Classic KGE methods include translational models (e.g, TransE (Bordes et al., 2013)) and semantic-matching models (e.g., ComplEx (Trouillon et al., 2016)). To enhance representation capability (Liang et al., 2024), hypercomplex spaces have been introduced: QuatE (Zhang et al., 2019) first extends embeddings to quaternion space, improving the modeling of symmetry and hierarchy. Subsequently, DualE (Cao et al., 2021) and BiQUE (Guo and Kok, 2021) further generalize to dual quaternions and biquaternion spaces, supporting richer relational composition via translation and rotation. Hypercomplex representations exhibit strong expressiveness for hierarchical, symmetric, and complex relational structures, and have recently been applied to more advanced KGC scenarios (Chung and Whang, 2023). However, prior hypercomplex-based methods focus only on unimodal knowledge graphs, and their potential for handling rich multi-modal semantics remains underexplored. In contrast, our approach is the first to leverage biquaternion space for MMKGs, supporting both multi-modality and complex relational transformations.

2.2 Multi-modal Knowledge Graph Completion

Existing Multi-modal Knowledge Graph Completion (MMKGC) methods extend traditional KGC models by integrating various modalities (e.g., structural information in MMKG, as well as image and textual information of entities). From the perspective of multi-modality modeling, current MMKGC methods can be categorized into multi-modal fusion-based methods and multi-modal ensemble-based methods.

Multi-modal fusion-based methods aim to design sophisticated multi-modal fusion modules to achieve modality alignment. Earlier modality fusion methods like IKRL (Xie et al., 2017) and

TransAE (Wang et al., 2019b) achieve efficient modality fusion by introducing cross-modal loss functions, demonstrating the effectiveness of cross-modal interactions. Furthermore, research community continues to propose more complex modality fusion designs with advanced techniques, such as OTKGE (Cao et al., 2022) with optimal transfer, AdaMF (Zhang et al., 2024) with adversarial training and MyGO (Zhang et al., 2025a) with fine-grained multi-modal tokenization. However, these modal fusion methods rarely preserve independent modalities and excessively rely on fixed fusion strategies. Therefore, this paradigm inevitably introduces information loss during the modality fusion stage and makes it difficult to adapt to the flexible modality requirements during the reasoning stage.

In contrast, classic modality ensemble methods like MoSE (Zhao et al., 2022) usually design individual sub-models for different modalities, and the individual representations obtained by these sub-models are integrated for joint decision-making. Subsequently, IMF (Li et al., 2023) utilizes tensor decomposition to fuse multi-modality information and introduces a sub-model of joint modalities into the modality ensemble method. We consider this a promising beginning for achieving joint decision-making that incorporates both fused and independent modalities. After that, MoMoK (Zhang et al., 2025b) follows this idea and decouples the modal representations through the MoE network with minimizing their mutual information. However, under the multi-modality ensemble paradigm, the sub-models lack explicit mechanisms for comprehensive cross-modal interaction, thereby limiting their overall modeling capability.

3 Preliminaries

Quaternion number system was first proposed by Hamilton (1844) to extend the complex numbers. The algebraic representation of a quaternion is typically expressed as:

$$Q = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}, \tag{1}$$

where the coefficient a is a real number representing real part, the coefficients b, c, d are real numbers representing imaginary part, and $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ are the orthogonal basis vectors or basis elements, which satisfy the following multiplication properties: $\mathbf{i}1 = 1\mathbf{i} = \mathbf{i}$, $\mathbf{j}1 = 1\mathbf{j} = \mathbf{j}$, $\mathbf{k}1 = 1\mathbf{k} = \mathbf{k}$, $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$, $\mathbf{ij} = -\mathbf{ji} = \mathbf{k}$, $\mathbf{jk} = -\mathbf{kj} = \mathbf{i}$, $\mathbf{ki} = -\mathbf{ik} = \mathbf{j}$, and $\mathbf{ijk} = -1$.

Hamilton Product can be regarded as "Quaternion Multiplication", which is composed of all standard multiplications of factors in quaternions, defined as:

$$\begin{aligned} Q_1 \otimes Q_2 = &(a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2) \\ &+ (a_1b_2 + b_1a_2 + c_1d_2 - d_1c_2)\mathbf{i} \\ &+ (a_1c_2 - b_1d_2 + c_1a_2 + d_1b_2)\mathbf{j} \\ &+ (a_1d_2 + b_1c_2 - c_1b_2 + d_1a_2)\mathbf{k}. \end{aligned} \tag{2}$$

Biquaternions further extend quaternions, and their algebra can be considered as a tensor product $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H}$, where $\mathbb{C}$ is the field of complex numbers and $\mathbb{H}$ is the division algebra of (real) quaternions. Biquaternions extend the coefficients of quaternions to complex numbers, denoted as:

$$Q = (a_r + a_i\mathbf{I}) + (b_r + b_i\mathbf{I})\mathbf{i} + (c_r + c_i\mathbf{I})\mathbf{j} + (d_r + d_i\mathbf{I})\mathbf{k}, \tag{3}$$

where $\mathbf{I}$ is the imaginary unit of the complex number field $\mathbb{C}$, satisfying $\mathbf{I}^2 = -1$. The algebra $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H}$ satisfies the commutation relations $\mathbf{Ii} = \mathbf{iI}$, $\mathbf{Ij} = \mathbf{jI}$, $\mathbf{Ik} = \mathbf{kI}$.

Hamilton Product of Biquaternions can be seen as an extension of the Hamilton product of quaternions. Similarly, for two biquaternions $Q_1 = a_1 + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k} = (a_{\mathrm{r},1} + a_{\mathrm{i},1}\mathbf{I}) + (b_{\mathrm{r},1} + b_{\mathrm{i},1}\mathbf{I})\mathbf{i} + (c_{\mathrm{r},1} + c_{\mathrm{i},1}\mathbf{I})\mathbf{j} + (d_{\mathrm{r},1} + d_{\mathrm{i},1}\mathbf{I})\mathbf{k}$ and $Q_2 = a_2 + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k} = (a_{\mathrm{r},2} + a_{\mathrm{i},2}\mathbf{I}) + (b_{\mathrm{r},2} + b_{\mathrm{i},2}\mathbf{I})\mathbf{i} + (c_{\mathrm{r},2} + c_{\mathrm{i},2}\mathbf{I})\mathbf{j} + (d_{\mathrm{r},2} + d_{\mathrm{i},2}\mathbf{I})\mathbf{k}$, the multiplication is performed exactly as in Equation 2 for quaternions, but with all coefficients treated as complex numbers (with $\mathbf{I}^2 = -1$). That is, the Hamilton product is defined in the same way, with addition and multiplication of coefficients carried out in the field of complex numbers $\mathbb{C}$.

4 Methodology

In this section, we introduce **M-Hyper**, which models **M**ulti-modal knowledge graphs (MMKG) in **Hyper**complex spaces. As shown in Figure 2, we utilize the Fine-grained Entity Representation Factorization (FERF) module and the Robust Relation-aware Modality Fusion (R2MF) module to obtain robust representations for three independent modalities and one fused modality. These modality representations are mapped to the four orthogonal bases of a biquaternion, enabling unified score modeling.

4.1 Problem Definition

A Multi-modal Knowledge Graph (MMKG) can be denoted as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}, \mathcal{R}$ are the

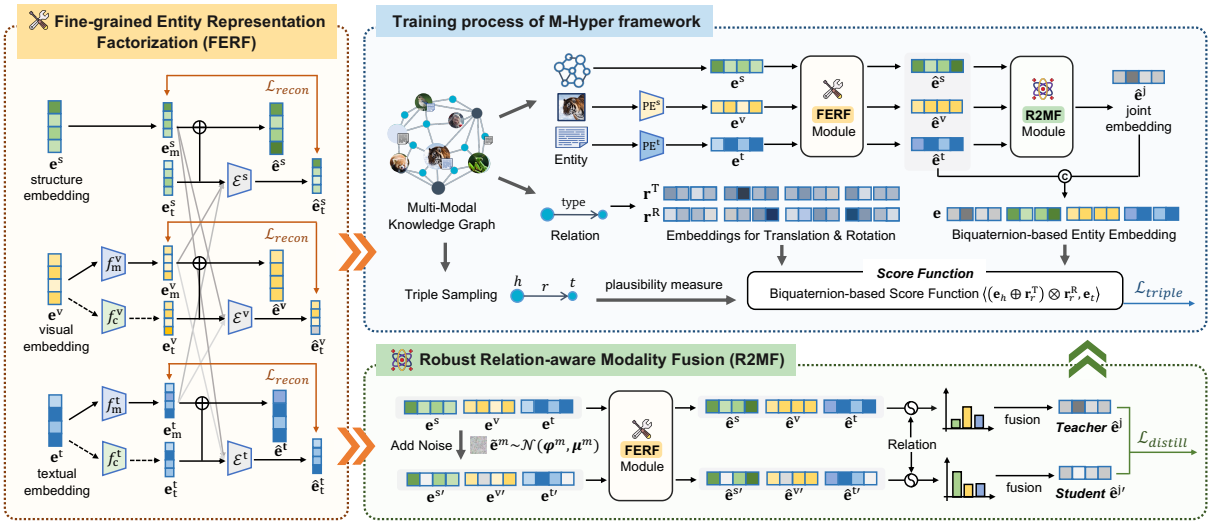

Figure 2: The overview of our M-Hyper, which integrates the Fine-grained Entity Representation Factorization (FERF) module and the Robust Relation-aware Modality Fusion (R2MF) module to learn robust representations for three modalities and their fusion, enabling unified multi-modal knowledge graph modeling in hypercomplex spaces.

entity set and relation set, and $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ represents the set of triples. Additionally, for each entity $e \in \mathcal{E}$, its modality information can be denoted as $\mathcal{X}^m(e)$ under a specific modality $m \in \mathcal{M}$. Specifically, $\mathcal{X}^m(e)$ can be a set of image or textual description for entity e, or structural information embodied in the KG $\mathcal{G}$.

Multi-modal Knowledge Graph Completion (MMKGC) models measure the plausibility of each triple $(h, r, t) \in \mathcal{T}$ using a score function ϕ to embed the entities and relations into a continuous vector space. We usually evaluate MMKGC models with the link prediction task, which requires predicting the missing head entity or tail entity for a given query $(?, r, t)$ or $(h, r, ?)$. For each candidate $e \in \mathcal{E}$, the score of the triple (h, r, e) or (e, r, t) is calculated and then ranked across the entire candidate entity set.

4.2 Fine-grained Entity Representation Factorization

Modality missing (Chen et al., 2023a) and cross-modal semantic ambiguity (Zhang et al., 2025a) consistently challenge the robustness of MMKGC models. These issues not only lead to information inconsistency across modalities but also introduce significant noise, making it more difficult to extract task-relevant semantic information, especially in scenarios requiring modality-cooperative reasoning. To address this problem, we decompose the representation of each individual modality m into two complementary semantic subspaces: (1) modality-specific representation $\mathbf{e}_{\mathrm{m}}^m \in \mathbb{R}^{2d}$ and (2) task-specific representation $\mathbf{e}_{\mathrm{t}}^m \in \mathbb{R}^{2d}$.

For modality-specific representation $\mathbf{e}_{\mathrm{m}}^m$, the structural embedding $\mathbf{e}_{\mathrm{m}}^{\mathrm{s}}$ is learned from scratch during training, while textual and visual modality embeddings are learned from the features extracted by the pre-trained model (Devlin et al., 2019; Simonyan and Zisserman, 2015), denoted as:

$$\mathbf{e}_{\mathrm{m}}^m = f_{\mathrm{m}}^m\Big(\frac{1}{|\mathcal{X}^m(e)|}\sum_{x^m \in \mathcal{X}^m(e)} \mathrm{PE}^m(x^m)\Big), \quad (4)$$

where $m \in \{\mathrm{t}, \mathrm{v}\}$, $f_{\mathrm{m}}^m : \mathbb{R}^{d^m} \to \mathbb{R}^{2d}$ is 1-layer MLP, $\mathcal{X}^m(e)$ is the set of modality information for m modality of entity e, and PE^m represents the pre-trained encoder. For task-specific representation $\mathbf{e}_{\mathrm{t}}^m$, they are all learnable embeddings during training. Among them, visual $\mathbf{e}_{\mathrm{t}}^{\mathrm{v}}$ and textual $\mathbf{e}_{\mathrm{t}}^{\mathrm{t}}$ embeddings are initialized by applying PCA to extract coarse-grained modal information from raw embeddings.

Furthermore, to ensure task-specific representations not only retain the unique characteristics of each independent modality but also enhance cross-modal collaborative representation capabilities, we introduce a reconstruction loss:

$$\mathcal{L}_{recon} = \sum_m ||\mathcal{E}^m(\mathbf{e}_{\mathrm{t}}^m; \{\mathbf{e}_{\mathrm{m}}^{\hat{m}} : \hat{m} \neq m\}) - \mathbf{e}_{\mathrm{m}}^m||^2, \quad (5)$$

where $\mathcal{E}^m$ is MLP. This loss requires the modality-specific embeddings to collaborate with other modalities to jointly reconstruct the original modality information. The final embedding is $\hat{\mathbf{e}}^m =$

$\mathbf{e}^m_{\mathrm{m}} + \mathbf{e}^m_{\mathrm{t}}$ for modality m, and entire module can be denoted as: $\hat{\mathbf{e}}^{\mathbf{s}}, \hat{\mathbf{e}}^{\mathrm{v}}, \hat{\mathbf{e}}^{\mathrm{t}} = \mathrm{FERF}(\mathbf{e}^{\mathbf{s}}, \mathbf{e}^{\mathrm{v}}, \mathbf{e}^{\mathrm{t}})$.

4.3 Robust Relation-aware Modality Fusion

In terms of relation representation, to model both translation and rotation transformations (Cao et al., 2021), we define Translation embeddings $\mathbf{r}^{\mathrm{T}} = ||^4_{i=1}\mathbf{r}^{\mathrm{T}}_i \in \mathbb{R}^{8d}$ and Rotation embeddings $\mathbf{r}^{\mathrm{R}} = ||^4_{i=1}\mathbf{r}^{\mathrm{R}}_i \in \mathbb{R}^{8d}$ for each relation r, and their algebraic representations are denoted as Q^{T}_r and Q^{R}_r.

Relation-aware Gated Fusion. Considering that modality information required by an entity varies across different relation queries, we aim to design a adaptive relation-aware fusion strategy for the entity's fused modality embeddings $\hat{\mathbf{e}}^{\mathrm{j}}$. Specifically, we first compute the contribution weights of the entity's modality embeddings under relation r:

$$w^m = f^m_w([\hat{\mathbf{e}}^m; \mathbf{r}^{\mathrm{T}}; \mathbf{r}^{\mathrm{R}}]), m \in \{\mathrm{s}, \mathrm{v}, \mathrm{t}\} \quad (6)$$

where $f^m_w : \mathbb{R}^{18d} \to \mathbb{R}^1$ are 1-layer MLPs. Then, when applying softmax to normalize the weights, we introduce a learnable relation-wise temperature coefficient τ_r to further optimize the weight distribution: $\hat{w}^m(e, r) = \frac{\exp(w^m/\tau_r)}{\sum_i \exp(w^i/\tau_r)}$. Consequently, during gated fusion process, we also equip entity e with a task-specific embedding $\mathbf{e}^{\mathrm{j}}_{\mathrm{t}} \in \mathbb{R}^{2d}$ denoted as:

$$\hat{\mathbf{e}}^{\mathrm{j}} = \sum_m \hat{w}^m \hat{\mathbf{e}}^m + \mathbf{e}^{\mathrm{j}}_{\mathrm{t}}, m \in \{\mathrm{s}, \mathrm{v}, \mathrm{t}\}. \quad (7)$$

Ultimately, we denote the entire relation-aware gated fusion process as: $\hat{\mathbf{e}}^{\mathrm{j}} = \mathrm{Rel}(\hat{\mathbf{e}}^{\mathbf{s}}, \hat{\mathbf{e}}^{\mathrm{v}}, \hat{\mathbf{e}}^{\mathrm{t}})$.

Noise-powered Self-distillation. Chen et al. (2023a) have found introducing a certain degree of modality noise into MMKGs can effectively enhance the robustness of the model's entity representations. Inspired by this, we aim to enhance the robustness of dynamic gated fusion by introducing modality noise. Specifically, given the original embedding set $\{\mathbf{e}^m_i\}^N_{i=1}$ of modality m, we can calculate the feature mean $\boldsymbol{\varphi}^m = \frac{1}{N}\sum^N_{i=1}\mathbf{e}^m_i$ and variance $\boldsymbol{\mu}^m = \frac{1}{N}\sum^N_{i=1}(\mathbf{e}^m_i - \boldsymbol{\varphi}_m)^2$. Next, we add Gaussian noise $\tilde{\mathbf{e}}^m \sim \mathcal{N}(\boldsymbol{\varphi}^m, \boldsymbol{\mu}^m)$ to a certain ratio β of original representations, denoted as: $\mathbf{e}^{\mathbf{s}\prime} = \mathbf{e}^{\mathrm{s}} + \tilde{\mathbf{e}}^m$. Furthermore, we take the fused embedding obtained without noise $\hat{\mathbf{e}}^{\mathrm{j}}_i$ as teacher and the fused embedding obtained with added noise $\hat{\mathbf{e}}^{\mathrm{j}\prime}_i$ as the student. During the training process, a self-distillation loss is introduced:

$$\mathcal{L}_{distill} = \frac{1}{n}\sum^n_{i=1}\left\|\hat{\mathbf{e}}^{\mathrm{j}}_i - \hat{\mathbf{e}}^{\mathrm{j}\prime}_i\right\|^2, \quad (8)$$

where $\hat{\mathbf{e}}^{\mathrm{j}}_i = \mathrm{Rel}(\mathrm{FERF}(\mathbf{e}^{\mathbf{s}}, \mathbf{e}^{\mathrm{v}}, \mathbf{e}^{\mathrm{t}}))$ is teacher embedding and $\hat{\mathbf{e}}^{\mathrm{j}\prime}_i = \mathrm{Rel}(\mathrm{FERF}(\mathbf{e}^{\mathbf{s}\prime}, \mathbf{e}^{\mathrm{v}\prime}, \mathbf{e}^{\mathrm{t}\prime}))$ is student embedding. They share parameters of Rel and FERF modules. Noise-powered perturbations enforce embedding consistency and enhance the fusion gate's noise robustness.

4.4 Training with Biquaternion-based Score Function

To enable the coexistence of one fused and three independent modalities, we represent them respectively as the real part and the three imaginary parts of a biquaternion for entity e. Its ***algebraic representation*** is: $Q = \hat{\mathbf{e}}^{\mathrm{j}} + \hat{\mathbf{e}}^{\mathrm{s}}\mathbf{i} + \hat{\mathbf{e}}^{\mathrm{v}}\mathbf{j} + \hat{\mathbf{e}}^{\mathrm{t}}\mathbf{k} = (\hat{\mathbf{e}}^{\mathrm{j}}_{\mathrm{r}} + \hat{\mathbf{e}}^{\mathrm{j}}_{\mathrm{i}}\mathbf{I}) + (\hat{\mathbf{e}}^{\mathrm{s}}_{\mathrm{r}} + \hat{\mathbf{e}}^{\mathrm{s}}_{\mathrm{i}}\mathbf{I})\mathbf{i} + (\hat{\mathbf{e}}^{\mathrm{v}}_{\mathrm{r}} + \hat{\mathbf{e}}^{\mathrm{v}}_{\mathrm{i}}\mathbf{I})\mathbf{j} + (\hat{\mathbf{e}}^{\mathrm{t}}_{\mathrm{r}} + \hat{\mathbf{e}}^{\mathrm{t}}_{\mathrm{i}}\mathbf{I})\mathbf{k}$, whose all coefficients $\hat{\mathbf{e}}^m = [\hat{\mathbf{e}}^m_{\mathrm{r}}; \hat{\mathbf{e}}^m_{\mathrm{i}}] \in \mathbb{R}^{2d}$ are parameterized as embeddings, and ***embedding representation*** is the concatenation of all coefficients, denoted as:

$$\begin{aligned}\mathbf{e} &= [\mathbf{e}^{\mathrm{j}}; \mathbf{e}^{\mathrm{s}}; \mathbf{e}^{\mathrm{v}}; \mathbf{e}^{\mathrm{t}}] \\ &= [\mathbf{e}^{\mathrm{j}}_{\mathrm{r}}; \mathbf{e}^{\mathrm{j}}_{\mathrm{i}}; \mathbf{e}^{\mathrm{s}}_{\mathrm{r}}; \mathbf{e}^{\mathrm{s}}_{\mathrm{i}}; \mathbf{e}^{\mathrm{v}}_{\mathrm{r}}; \mathbf{e}^{\mathrm{v}}_{\mathrm{i}}; \mathbf{e}^{\mathrm{t}}_{\mathrm{r}}; \mathbf{e}^{\mathrm{t}}_{\mathrm{i}}] \in \mathbb{R}^{8d}\end{aligned} \quad (9)$$

Considering that the imaginary units of the quaternion field $\mathbb{H}$ are symmetric, the permutation of these modalities does not affect the representation.

Biquaternion-based Score Function. We adopt a standard semantic-matching (Liang et al., 2024; Guo and Kok, 2021) strategy to score the plausibility of triple. For a given triple (h, r, t), we first apply the following algebraic operation to calculate the embedding of query $(h, r, ?)$:

$$Q_{h''} = (Q_h \oplus Q^{\mathrm{T}}_r) \otimes Q^{\mathrm{R}}_r, \quad (10)$$

where $\oplus$ and $\otimes$ represent addition and Hamilton product between biquaternions. The addition is an element-wise sum: $Q_{h'} = Q_h \oplus Q^{\mathrm{T}}_r = (\mathbf{e}^{\mathrm{j}}_h + \mathbf{r}^{\mathrm{T}}_1) + (\mathbf{e}^{\mathbf{s}}_h + \mathbf{r}^{\mathrm{T}}_2)\mathbf{i} + (\mathbf{e}^{\mathrm{v}}_h + \mathbf{r}^{\mathrm{T}}_3)\mathbf{j} + (\mathbf{e}^{\mathrm{t}}_h + \mathbf{r}^{\mathrm{T}}_4)\mathbf{k} = \mathbf{e}^{\mathrm{j}}_{h'} + \mathbf{e}^{\mathbf{s}}_{h'}\mathbf{i} + \mathbf{e}^{\mathrm{v}}_{h'}\mathbf{j} + \mathbf{e}^{\mathrm{t}}_{h'}\mathbf{k}$ for characterizing translation transformations. Then Q^{T}_r rotates the query via Hamilton product as shown in Equation 2:

$$Q_{h''} = Q_{h'} \otimes Q^{\mathrm{R}}_r = \sum_m^{|\mathcal{M}|}\sum_{k=1}^{4} H_{ikm}(\mathbf{e}^m_{h'}) \circledast (\mathbf{r}^{\mathrm{R}}_k)\,\mathbf{u}_i, \quad (11)$$

where H_{ikm} denotes the structure constants uniquely defining the multiplication rules for biquaternion algebra, and $\mathbf{u}_i \in \{1, \mathbf{i}, \mathbf{j}, \mathbf{k}\}$ indicates the corresponding basis element, and $\circledast$ represents multiplication in complex number field $\mathbb{C}$.

Figure 3: Compared to existing MMKGC score functions, M-Hyper achieves the most comprehensive modality interaction and geometric transformation. For detailed formulaic theoretical proofs, please refer to Appendix A.

Optimization Objective. We utilize the standard vector dot-product between query Q''_h and tail entity $Q_t = \mathbf{e}_t^{\mathrm{j}} + \mathbf{e}_t^{\mathrm{s}}\mathbf{i} + \mathbf{e}_t^{\mathrm{v}}\mathbf{j} + \mathbf{e}_t^{\mathrm{t}}\mathbf{k}$ to compute plausibility score: $\phi(h,r,t) = \langle Q_{h''}, Q_t \rangle = [\mathbf{e}_{h''}^{\mathrm{j}}; \mathbf{e}_{h''}^{\mathrm{s}}; \mathbf{e}_{h''}^{\mathrm{v}}; \mathbf{e}_{h''}^{\mathrm{t}}] \cdot [\mathbf{e}_t^{\mathrm{j}}; \mathbf{e}_t^{\mathrm{s}}; \mathbf{e}_t^{\mathrm{v}}; \mathbf{e}_t^{\mathrm{t}}]^\top$. We optimize our model using the cross-entropy loss:

$$\mathcal{L}_{triple} = \sum_{t'}^{|\mathcal{V}|} \log(1 + \exp(y_{t'}\phi(h,r,t')), \quad (12)$$

where $y_{t'}$ is the ground-truth of the candidate tail entity t'. So the Biquaternion-based score function can be expressed as: $\phi(h,r,t) = \langle (Q_h \oplus Q_r^{\mathrm{T}}) \otimes Q_r^{\mathrm{R}}, Q_t \rangle$. As shown in Figure 3, it can achieve the most comprehensive modal interaction and geometric transformation (translation + rotation). We provide a more in-depth theoretical proof in Appendix A. The overall training objective $\mathcal{L}_{total}$ is represented as:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \mathcal{L}_{distill} + \mathcal{L}_{triple} + \mathcal{L}_{reg}, \quad (13)$$

where we also employ N3 regularization norm (Lacroix et al., 2018) to prevent overfitting: $\mathcal{L}_{reg} = \lambda(||\mathbf{e}_h||_3^3 + ||\mathbf{r}_r^{\mathrm{T}}||_3^3 + ||\mathbf{r}_r^{\mathrm{R}}||_3^3 + ||\mathbf{e}_t||_3^3)$, and λ is a regularization hyperparameter.

5 Experiments

5.1 Experimental Settings

Datasets. The experiments are conducted on three common MMKG benchmarks: DB15K (Liu et al., 2019), MKG-W (Xu et al., 2022) and MKG-Y (Xu et al., 2022). To ensure fairness in comparison with previous works, we adopt the same representations of the visual and textual modalities in the original datasets derived from the pre-trained models VGG (Simonyan and Zisserman, 2015) and BERT (Devlin et al., 2019). DB15K (Liu et al., 2019) is a subset of DBPedia (Lehmann et al., 2015) with images crawled from search engines. MKG-W and MKG-Y (Xu et al., 2022) are derived from Wikidata (Vrandecic and Krötzsch, 2014) and YAGO (Suchanek et al., 2007) respectively. The detailed statistics are shown in Appendix F.

Evaluation Protocols. Link prediction tasks need to predict the missing entity of a given query $(h,r,?)$ or $(?,r,t)$ from $\mathcal{T}_{test}$. Consistent with the existing works, We use Mean Reciprocal Rank (MRR) and Hit@K (K=1, 3, 10) to evaluate the results. MRR and Hit@K metrics can be calculated as: $\mathbf{MRR} = \frac{1}{|\mathcal{T}_{test}|}\sum_{i=1}^{|\mathcal{T}_{test}|}(\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}})$, $\mathbf{Hit@K} = \frac{1}{|\mathcal{T}_{test}|}\sum_{i=1}^{|\mathcal{T}_{test}|}(\mathbf{1}(r_{h,i} \leq K) + \mathbf{1}(r_{t,i} \leq K))$, where $r_{h,i}$ and $r_{t,i}$ are the results of head prediction and tail prediction respectively. Besides, we apply filter setting (Bordes et al., 2013) to eliminate existing facts in the dataset.

Baselines. We select 19 representative MMKGC methods as our baselines, including: **(1) Uni-modal KGC methods**: TransE (Bordes et al., 2013), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), QuatE (Zhang et al., 2019), DualE (Cao et al., 2021) ,and BiQUE (Guo and Kok, 2021). These methods only model structural information of the KGs. **(2) Multi-modal KGC models**: *fusion-based* methods: IKRL (Xie et al., 2017), TransAE (Wang et al., 2019b), VBKGC (Zhang and Zhang, 2022), OTKGE (Cao et al., 2022), QEB (Wang et al., 2023), VISTA (Lee et al., 2023), AdaMF (Zhang et al., 2024), MyGO (Zhang et al., 2025a), K-ON (Guo et al., 2025), *ensemble-based* methods: MoSE (Zhao et al., 2022), IMF (Li et al., 2023), MoMoK (Zhang et al., 2025b). These methods utilize both the structural information and multi-modal information in the KGs, among which K-ON (Guo et al., 2025) is the most advanced LLM-based method.

Implementation Details. All experiments are conducted on a Nvidia A800 GPU and implemented with PyTorch. We also add inverse triple (t, r^{-1}, h) for each observed triple (h,r,t) in train-set as training samples. We use Adagrad (Duchi et al., 2011) as the optimizer. For hyperparameters, batch size is fixed at 1000; and we search the learning rate $\alpha \in \{\mathbf{0.1}, 0.05, 0.01, 0.005\}$; dimension of embeddings $d \in \{64, \mathbf{128}, 256\}$; regularization factors $\lambda \in \{0.01, \mathbf{0.005}, 0.001\}$ and noise rate

Model		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
Uni-modal KGC	TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
	ComplEx	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73	28.71	22.26	32.12	40.93
	RotatE	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
	QuatE*	34.18	25.42	38.91	51.30	34.50	28.94	36.71	46.64	36.01	30.53	38.84	43.68
	DualE*	35.85	29.31	38.52	51.28	33.94	27.55	36.56	46.09	34.95	29.77	38.44	43.12
	BiQUE*	38.34	32.38	41.48	53.23	35.01	29.42	37.01	46.49	36.74	34.82	38.25	42.16
Multi-modal KGC	IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26
	TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
	VBKGC	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	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38
	MoSE	28.38	21.56	30.91	41.67	33.34	27.78	33.94	41.06	36.28	33.64	37.47	40.81
	MMRNS	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47
	QEB	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32	34.37	29.49	36.95	42.32
	VISTA	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61	30.45	24.87	32.39	41.53
	IMF	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44	35.79	32.95	37.14	40.63
	AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.21	38.06	33.49	40.44	**45.48**
	MyGO	37.72	30.08	41.26	52.21	36.10	29.78	38.54	47.75	38.44	35.01	39.84	44.19
	K-ON*	36.24	28.13	40.49	51.26	35.83	29.41	37.32	47.16	35.83	32.56	37.34	42.45
	MoMoK	39.57	32.38	43.45	54.14	35.89	30.38	37.54	46.13	37.91	35.09	39.20	43.20
Ours	**M-Hyper**	**41.25**	**33.64**	**45.01**	**56.09**	**37.02**	**31.24**	**39.16**	**48.84**	**39.46**	**36.02**	**40.92**	45.22

Table 1: Results on DB15K, MKG-W, and MKG-Y datasets. The best results are marked **bold** and the second-best results are underlined. The *results are reproduced by us, and others are taken from MoMoK (Zhang et al., 2025b).

$\beta \in \{0.1, \mathbf{0.2}, 0.4\}$.

5.2 Main Results

The experimental results are shown in Table 1. M-Hyper outperforms 18 existing baselines on most metrics, including AdaMF and MoMoK, which also adopt modality noise enhancement. Specifically, M-Hyper achieves a 4.25% improvement in MRR and a 3.89% improvement in Hit@10, demonstrating significant performance improvements. Compared to the classic fusion-based (Li et al., 2023) and ensemble-based (Zhang et al., 2025a) paradigm, M-Hyper not only preserves the original modality information but also enables dynamic and flexible modality interaction, providing a promising modeling paradigm for MMKGC task.

5.3 Efficiency Analysis

We conduct an efficiency analysis of M-Hyper focusing on memory usage and runtime, with the results shown in Figure 4(a). Compared to 6 state-of-the-art methods, our M-Hyper achieves the best training efficiency, requiring the shortest runtime for a single training epoch. In terms of memory usage, M-Hyper demonstrates nearly optimal performance. Figure 4(b) illustrates the training times required to achieve the best performances. Our model requires only 1160 seconds of training time to achieve an MRR of 40.75% and a Hit@1 of 33.14%. So we can conclude M-Hyper not only delivers the best performance but also achieves the highest computational efficiency with the least memory usage and the shortest training time.

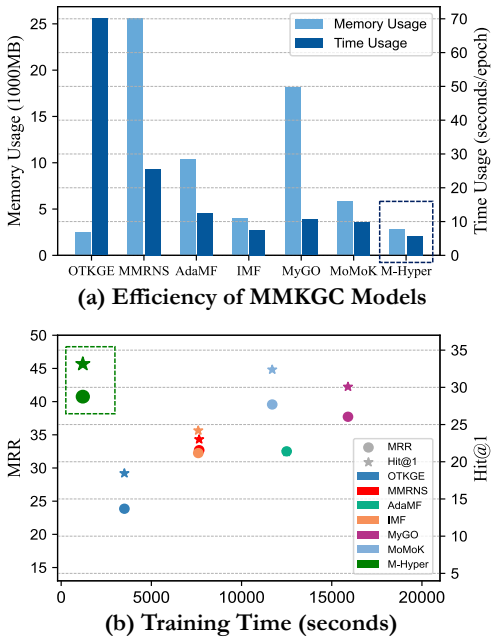

Figure 4: Efficiency results on memory usage, training time usage, and the trade-off between training effectiveness and training time on DB15K dataset.

5.4 Ablation Study

Modality Ablation Study. To verify the contributions of each modality, we set the corresponding modality embedding to an all-zero embedding, removing its influence. As shown in Table 2, all modalities positively impact performance, albeit

	Setting	DB15K		MKG-W		MKG-Y	
		MRR	Hit@1	MRR	Hit@1	MRR	Hit@1
$\mathcal{G}_0$	w/o joint $\hat{\mathbf{e}}^{\mathrm{j}}$	36.36	28.54	35.09	29.16	36.71	33.42
	w/o structure $\hat{\mathbf{e}}^{\mathrm{s}}$	39.77	32.17	34.62	28.63	38.03	34.60
	w/o vision $\hat{\mathbf{e}}^{\mathrm{v}}$	35.09	27.22	36.46	30.60	37.95	34.68
	w/o text $\hat{\mathbf{e}}^{\mathrm{t}}$	39.70	32.12	36.28	31.17	38.09	34.74
$\mathcal{G}_1$	w/o FERF	39.24	31.83	35.93	29.38	37.93	34.53
	w/o noise-powered	39.64	32.16	36.10	30.28	38.16	35.82
	w/o r-aware gate	40.18	32.47	36.18	30.44	38.21	35.14
	w/o $\mathcal{L}_{recon}$	40.97	33.24	36.18	30.69	39.12	35.23
	w/o translation $\mathbf{r}^{\mathrm{T}}$	39.50	31.42	35.13	29.56	37.86	34.64
	w/o rotation $\mathbf{r}^{\mathrm{R}}$	38.91	31.35	36.46	30.67	37.78	34.55
	M-Hyper+DualE	39.93	32.07	35.96	30.10	38.02	34.78
$\mathcal{G}$	M-Hyper-fusion	39.23	31.66	35.54	30.35	37.52	34.51
	M-Hyper-ensemble	39.31	31.71	34.75	29.26	37.58	34.78
	M-Hyper(ours)	**41.25**	**33.64**	**37.02**	**31.24**	**39.46**	**36.02**

Table 2: Results of modality ablation $\mathcal{G}_0$ and model ablation $\mathcal{G}_1$. $\mathcal{G}$ represents the comparison among the three modality modeling paradigms.

to varying degrees across different datasets. Notably, excluding the joint modality leads to the most substantial performance decline, highlighting its pivotal role in M-Hyper's overall effectiveness.

Model Ablation Study. We can see that each module contributes to the overall performance. FERF and noise-powered self-distillation modules enable more robust modality representations, while the relation-aware gate facilitates dynamic modality fusion to handle complex contexts. Additionally, translation and rotation relation embeddings enable more sophisticated relational modeling. Notably, removing the rotation operation $\mathbf{r}^{\mathrm{R}}$ in complex field $\mathbb{C}$ reduces hypercomplex space to a quaternion space and results in a performance decline, indicating that the biquaternion space offers greater expressive power. Meanwhile, we introduce M-Hyper variants under the ensemble and fusion paradigms, whose score functions are provided in Appendix C. It can be observed that M-Hyper, benefiting from adequate collaboration between independent and fused modalities, achieves the best performance.

5.5 Robustness to Complex Scenarios

Following Zhang et al. (2025b), we evaluate MMKGC robustness under three challenging scenarios: (1) modality missing, (2) modality noise, and (3) link sparsity. To be specific, in modality missing scenario, we randomly delete a certain ratio of entity's raw modality embeddings. For the modality noise scenario, we randomly add Gaussian noise to raw modality embeddings. In the link sparsity scenario, we randomly remove a certain ratio of training triples.

As shown in Figure 5, the model's performance declines to varying degrees under these complex

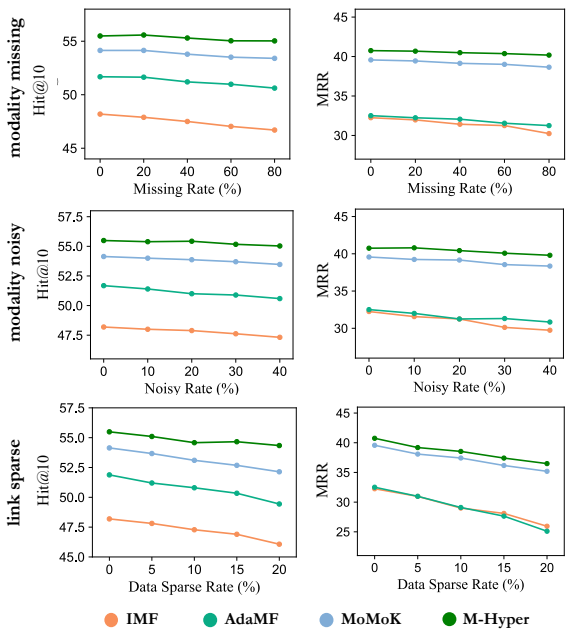

Figure 5: Results on DB15K under 3 complex scenarios: modality missing, modality noisy and link sparse.

scenarios. Among them, the training data, as a critical source of structural information, significantly contributes to the model's performance. Notably, we find that AdaMF, MoMoK, and M-Hyper with noise-augmented training achieve improved robustness. Moreover, unlike previous noise-augmented methods, we introduce task-specific representations and a self-distillation supervision strategy, which further enhance model's noise-reduction capabilities and improve the effectiveness of dynamic fusion. As a result, our approach achieves relatively superior robust performance.

5.6 Modality Visualization Analysis

As illustrated in Figure 6, we apply t-SNE to visualize the modality embeddings of cities across 6 countries in DB15K dataset. It is evident that the presence of modality ambiguity and bias introduces variability in the representation efficacy of entities across different modalities. Notably, among all modalities, the joint modality representations demonstrate the highest discriminative capability in differentiating entities. Furthermore, the integration of the FERF and R2MF modules significantly improves the expressiveness and effectiveness of the modality-specific embeddings, highlighting their ability to mitigate modality bias and enhance representation quality.

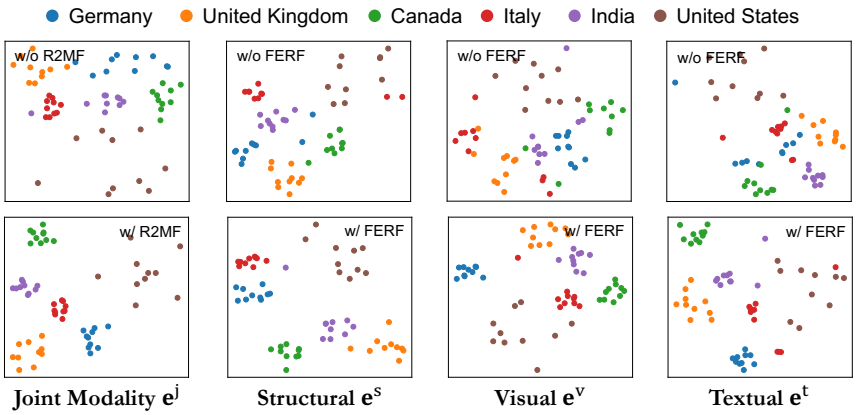

Figure 6: Embedding visualization under t-SNE for cities under relation *country*, and distinct colors are utilized to represent different countries.

6 Conclusion

In this paper, we highlight the limitations of existing MMKGC paradigms, which struggle to balance fused and independent modality representations. To enable efficient and flexible cross-modal collaboration, we propose M-Hyper, the first method to represent MMKGs in hypercomplex space. Specifically, we introduce Fine-grained Entity Representation Factorization (FERF) module and Robust Relation-aware Modality Fusion (R2MF) module to obtain robust representations for three independent modalities and one fused modality. Subsequently, these modality representations are mapped onto the four orthogonal bases of a biquaternion, enabling efficient modeling of pairwise interactions and comprehensive cross-modal integration. Empirical results show that our M-Hyper demonstrate greater performance and robustness.

Limitations

We focus on "transductive" multi-modal knowledge graph completion (MMKGC) under a static setting, assuming that entities, relations, and modality information remain fixed during both training and inference. Therefore, for dynamic scenarios with entities, relations, or modality features (e.g., newly added images or textual descriptions) undergoing frequent updates, it may be necessary to design online learning frameworks or dynamic modeling approaches to address evolving data distributions and incremental modality adaptation. In addition, we also hope to explore the idea of coexistence of independence and integration in other task scenarios, such as entity alignment (Xiao et al., 2025), named entity recognition (Pang et al., 2024), and knowledge graph question answering (Gong et al., 2026).

Ethics Statement

In this paper, we explore the multi-modal knowledge graph completion task with deep learning techniques. Our training and evaluation are based on publicly available and widely used datasets of different types of knowledge graphs. Therefore, we believe this does not violate any ethics.

Acknowledgments

This work is founded by National Natural Science Foundation of China (NSFCU23B2055/NSFC62306276), New Generation Artificial Intelligence-National Science and Technology Major Project 2030 (2025ZD0122800), Yongjiang Talent Introduction Programme (2022A-238-G), and Fundamental Research Funds for the Central Universities (226-2023-00138). This work was supported by Ant Group.

References

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. 2021. Dual quaternion knowledge graph embeddings. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh*

Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 6894–6902. AAAI Press.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: multi-modal knowledge graph embeddings via optimal transport. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z. Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023a. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I*, volume 14265 of *Lecture Notes in Computer Science*, pages 121–139. Springer.

Zhuo Chen, Wen Zhang, Yufeng Huang, Mingyang Chen, Yuxia Geng, Hongtao Yu, Zhen Bi, Yichi Zhang, Zhen Yao, Wenting Song, Xinliang Wu, Yi Yang, Mingyi Chen, Zhaoyang Lian, Yingying Li, Lei Cheng, and Huajun Chen. 2023b. Tele-knowledge pre-training for fault analysis. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, pages 3453–3466. IEEE.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *CoRR*, abs/2402.05391.

Chanyoung Chung and Joyce Jiyoung Whang. 2023. Learning representations of bi-level knowledge graphs for reasoning beyond link prediction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 4208–4216. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159.

Zhaoyan Gong, Zhiqiang Liu, Songze Li, Xiaoke Guo, Yuanxiang Liu, Xinle Deng, Zhizhen Liu, Lei Liang, Huajun Chen, and Wen Zhang. 2026. Temp-r1: A unified autonomous agent for complex temporal kgqa via reverse curriculum reinforcement learning. *arXiv preprint arXiv:2601.18296*.

Jia Guo and Stanley Kok. 2021. Bique: Biquaternionic embeddings of knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8338–8351. Association for Computational Linguistics.

Lingbing Guo, Yichi Zhang, Zhongpu Bo, Zhuo Chen, Mengshu Sun, Zhiqiang Zhang, Yangyifei Luo, Wen Zhang, and Huajun Chen. 2025. K-on: Knowledge on the head layer of large language model. In *AAAI*.

William Rowan Hamilton. 1844. Lxxviii. on quaternions; or on a new system of imaginaries in algebra: To the editors of the philosophical magazine and journal. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 25(169):489–495.

Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical tensor decomposition for knowledge base completion. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2869–2878. PMLR.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. 2023. VISTA: visual-textual knowledge graph representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7314–7328. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2572–2580. ACM.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. 2024. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9456–9478.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, volume 11503 of *Lecture Notes in Computer Science*, pages 459–474. Springer.

Zhiqiang Liu, Chengtao Gan, Junjie Wang, Yichi Zhang, Zhongpu Bo, Mengshu Sun, Huajun Chen, and Wen Zhang. 2025. Ontotune: Ontology-driven self-training for aligning large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 119–133.

Zhiqiang Liu, Yin Hua, Mingyang Chen, Zhuo Chen, Ziqi Liu, Lei Liang, Huajun Chen, and Wen Zhang. 2024. Unihr: Hierarchical representation learning for unified knowledge graph link prediction. *arXiv preprint arXiv:2411.07019*.

Jinhui Pang, Xinyun Yang, Xiaoyao Qiu, Zixuan Wang, and Taisheng Huang. 2024. Mmaf: Masked multimodal attention fusion to reduce bias of visual features for named entity recognition. *DATA INTELLIGENCE*, 6(4):1114–1133.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*, pages 1–5. IEEE.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019a. KGAT: knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 950–958. ACM.

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 2391–2399. ACM.

Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019b. Multimodal data enhanced representation learning for knowledge graphs. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, pages 1–8. IEEE.

Peng Xiao, Chao Liu, Wei Jia, and Lijun Dong. 2025. Aligned-entities-based fusion embedding on hetero-field knowledge graphs. *DATA INTELLIGENCE*, 7(3):618–635.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3140–3146. ijcai.org.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3857–3866. ACM.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. 2019. Quaternion knowledge graph embeddings. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2731–2741.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2025a. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 13322–13330. AAAI Press.

Yichi Zhang, Zhuo Chen, Lingbing Guo, yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2025b. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. In *The Thirteenth International Conference on Learning Representations*.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 2024. Unleashing the power of

imbalanced modality information for multi-modal knowledge graph completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 17120–17130. ELRA and ICCL.

Yichi Zhang and Wen Zhang. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *CoRR*, abs/2209.07084.

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. 2022. Mose: Modality split and ensemble for multimodal knowledge graph completion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10527–10536. Association for Computational Linguistics.

Appendix

A Detailed Proof of Theorem

Theorem 1: Let $X = \{M_{\text{s}}, M_{\text{v}}, M_{\text{t}}\}$ represent the multi-modal input and Y the target task. The M-Hyper representation is defined as: $Q = T_{\text{j}}\mathbf{1} + T_{\text{s}}\mathbf{i} + T_{\text{v}}\mathbf{j} + T_{\text{t}}\mathbf{k}$, where T_{j} encodes fused information across modalities, and $T_{\text{s}}, T_{\text{v}}, T_{\text{t}}$ preserve modality-specific information. Under the Information Bottleneck (IB) (Tishby and Zaslavsky, 2015) framework, with the IB loss:

$$\mathcal{L}_{\text{IB}}(T) = I(X;T) - \beta I(T;Y),$$

the M-Hyper representation achieves a strictly lower IB loss:

$$\mathcal{L}_{\text{IB}}(Q) < \min\left(\mathcal{L}_{\text{IB}}(T_f),\ \mathcal{L}_{\text{IB}}(T_{\text{ens}})\right), \tag{14}$$

where T_f is the fused representation and T_{ens} the ensemble representation.

Proof 1: Consider three representations: (1) M-Hyper Q, (2) fusion-based $T_f = f(X)$, and (3) ensemble $T_{\text{ens}} = \{T_{\text{j}}, T_{\text{s}}, T_{\text{v}}, T_{\text{t}}\}$. On the one hand, fusion T_f over-compresses and includes redundancy:

$$I(X;T_f) - I(X;Q) = \Delta_{\text{redundancy}} \tag{15}$$

$$= \sum_{i\neq j} I(T_i;T_j|Y) > 0, \tag{16}$$

where $\Delta_{\text{redundancy}}$ measures cross-modal redundancy that does not contribute to Y. On the other hand, ensemble T_{ens} lacks explicit interactions:

$$I(T_{\text{ens}};Y) \leq \tag{17}$$

$$\sum_i I(T_i;Y) + I(T_{\text{fuse}};Y) - \sum_{i<j} I(T_i;T_j;Y), \tag{18}$$

where triple mutual information $\sum_{i<j} I(T_i;T_j;Y)$ captures cross-modal synergy not fully utilized in a simple ensemble. Our Q in quaternion space $\mathbb{H}$ (via Hamilton product, see Theorem 1) generates interaction terms $C_{ij} = T_i \cdot T_j$ that satisfy:

$$\sum_{i<j} I(C_{ij};Y) \geq \eta \left\| T_i^\top T_j \right\|^2 > 0,$$

i.e., these interactions are informative for predicting Y. Imposing orthogonality ($\langle T_i, T_j\rangle = 0,\ i \neq j$) further reduces intra-representation redundancy, so

$$I(X;Q) < I(X;T_{\text{ens}}).$$

As Q contains all modality-specific information (from $T_{\text{j}}, T_{\text{s}}, T_{\text{v}}, T_{\text{t}}$) plus explicit cross-modal interactions (i.e., C_{ij}), it is at least as informative about Y as T_{ens}, and typically more so:

$$I(Q;Y) \geq I(T_{\text{ens}};Y).$$

Combining the above, the difference in IB loss between Q and T_{ens} becomes

$$\begin{aligned}&\mathcal{L}_{\text{IB}}(Q) - \mathcal{L}_{\text{IB}}(T_{\text{ens}}) \\ &= \left[I(X;Q) - I(X;T_{\text{ens}})\right] \\ &\quad - \beta\left[I(Q;Y) - I(T_{\text{ens}};Y)\right] < 0.\end{aligned} \tag{19}$$

The first term is negative (due to reduced redundancy), and the second term is non-positive (due to improved relevance); thus their sum is strictly negative under $\beta > 0$. The comparison with T_f is similar, as detailed before:

$$\mathcal{L}_{\text{IB}}(Q) - \mathcal{L}_{\text{IB}}(T_f) \leq -\Delta_{\text{redundancy}} - \beta\Delta_{\text{interaction}} < 0,$$

where $\Delta_{\text{interaction}} = I(Q;Y) - I(T_f;Y) \geq 0$. So we can conclude:

$$\mathcal{L}_{\text{IB}}(Q) < \min\left(\mathcal{L}_{\text{IB}}(T_f),\ \mathcal{L}_{\text{IB}}(T_{\text{ens}})\right) \tag{20}$$

Therefore, Q achieves a strictly lower IB loss by both reducing redundancy (better compression of X) and boosting task relevance (enhanced dependence on Y) via explicit cross-modal interactions.

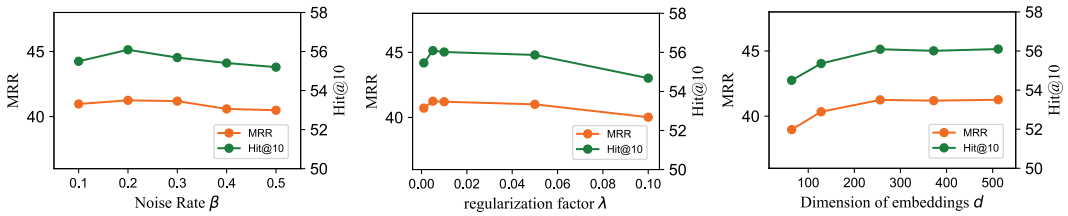

Figure 7: Results of hyperparamter analysis for noise rate β, regularization factor λ and dimension d.

Theorem 2: Let the entity embedding be $Q = \mathbf{e}^{\mathrm{j}}\mathbf{1} + \mathbf{e}^{\mathrm{s}}\mathbf{i} + \mathbf{e}^{\mathrm{j}}\mathbf{j} + \mathbf{e}^{\mathrm{t}}\mathbf{k}$, and score function as:

$$\phi(h, r, t) = \langle (Q_h \oplus Q_r^{\mathrm{T}}) \otimes Q_r^{\mathrm{R}}, Q_t \rangle \tag{21}$$

Then for any modalities $m, m' \in \{\mathrm{j}, \mathrm{s}, \mathrm{v}, \mathrm{t}\}$, the algebraic expansion contains all pair-wise interaction $\mathbf{e}_h^m \cdot \mathbf{e}_t^{m'}$, as well as intra-modal terms.

Proof 2: For the sake of simplicity in representation, we mark the final representations of each modality m as: $\mathbf{e}^m$. Our biquaternion-based score function can be expanded as:

$$\begin{aligned}
&\phi(h,r,t) = \langle (Q_h \oplus Q_r^{\mathrm{T}}) \otimes Q_r^{\mathrm{R}}, Q_t \rangle \\
&= \langle [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) + (\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}})\mathbf{i} + (\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,3}^{\mathrm{T}})\mathbf{j} + (\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,4}^{\mathrm{T}})\mathbf{k}] \\
&\quad \otimes [\mathbf{r}_{r,1}^{\mathrm{R}} + \mathbf{r}_{r,2}^{\mathrm{R}}\mathbf{i} + \mathbf{r}_{r,3}^{\mathrm{R}}\mathbf{j} + \mathbf{r}_{r,4}^{\mathrm{R}}\mathbf{k}], [\mathbf{e}_t^{\mathrm{j}} + \mathbf{e}_t^{\mathrm{s}}\mathbf{i} + \mathbf{e}_t^{\mathrm{v}}\mathbf{j} + \mathbf{e}_t^{\mathrm{t}}\mathbf{k}] \rangle \\
&= \langle [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}} - (\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}} - \\
&\quad (\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}} - (\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}}], \mathbf{e}_t^{\mathrm{j}} \rangle \\
&+ \langle [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}} + (\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}} + \\
&\quad (\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}} - (\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}}], \mathbf{e}_t^{\mathrm{s}} \rangle \\
&+ \langle [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}} - (\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}} + \\
&\quad (\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}} + (\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}}], \mathbf{e}_t^{\mathrm{t}} \rangle \\
&+ \langle [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}} + (\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}} - \\
&\quad (\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}} + (\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}}], \mathbf{e}_t^{\mathrm{v}} \rangle \\
&= [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{j}})^{\top} - [(\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{j}})^{\top} - \\
&\quad [(\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{j}})^{\top} - [(\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{j}})^{\top} \\
&+ [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{s}})^{\top} + [(\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{s}})^{\top} + \\
&\quad [(\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{s}})^{\top} - [(\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{s}})^{\top} \\
&+ [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{t}})^{\top} - [(\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{t}})^{\top} + \\
&\quad [(\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{t}})^{\top} + [(\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{t}})^{\top} \\
&+ [(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_{r,1}^{\mathrm{T}}) \circledast \mathbf{e}_{r,4}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{v}})^{\top} + [(\mathbf{e}_h^{\mathrm{s}} + \mathbf{r}_{r,2}^{\mathrm{T}}) \circledast \mathbf{e}_{r,1}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{v}})^{\top} - \\
&\quad [(\mathbf{e}_h^{\mathrm{t}} + \mathbf{r}_{r,3}^{\mathrm{T}}) \circledast \mathbf{e}_{r,2}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{v}})^{\top} + [(\mathbf{e}_h^{\mathrm{v}} + \mathbf{r}_{r,4}^{\mathrm{T}}) \circledast \mathbf{e}_{r,3}^{\mathrm{R}}] \cdot (\mathbf{e}_t^{\mathrm{v}})^{\top} \\
&= \sum_{m}^{|\mathcal{M}|} \sum_{m'}^{|\mathcal{M}|} \langle \mathcal{R}_{imm'}(\mathbf{e}_h^m), \mathbf{e}_t^{m'} \rangle.
\end{aligned} \tag{22}$$

where $\mathcal{R}_{imm'}$ represents the biquaternion algebra translation and rotation transformation between modality m and m', as intuitively shown in Figure 3. Based on the expanded formulation above, we observe that the biquaternion-based score function can be expressed as a linear combination of all pair-wise modality-specific score functions. Furthermore, these score functions independently characterize the translation and rotation of relationships. As a result, M-Hyper is capable of capturing all pairwise semantic relationships $m \to m'$, ensuring no information redundancy or missing modality combinations.

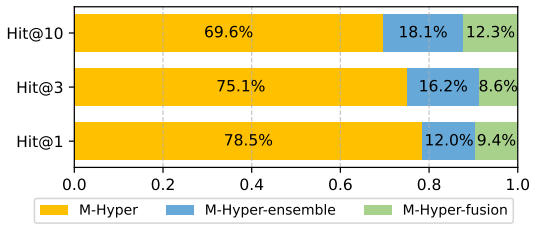

Figure 8: Comparison of Paradigm Proportions Achieving Optimal Performance Across Relations.

Relation	#Num	IMF	AdaMF	MyGO	MoMoK	M-Hyper
type‡	209	37.62	34.53	50.68	40.91	**52.26**
country‡	352	34.97	34.42	**48.97**	39.16	47.34
language‡	82	39.69	35.75	45.99	42.22	**50.73**
time_Zone‡	125	34.21	35.08	**59.71**	37.98	54.22
spouse◇	48	34.52	28.12	55.85	36.23	**65.05**
different_From◇	43	23.20	32.21	38.04	30.98	**40.67**
is_Part_Of†	183	32.98	30.05	72.16	39.44	**72.77**
company§	26	29.62	34.46	67.15	39.44	**83.19**
music_Composer†	151	37.61	32.23	47.62	42.20	**55.69**
associated_Band§	255	40.67	33.09	80.76	43.42	**87.17**

Table 3: Results of MRR per relation on DB15K. We mark 1-to-N†, N-to-1‡, N-to-N§, and symmetric◇ relations.

B Hyperparameter Analysis

We conducted an analysis of the hyperparameters involved in M-Hyper, with the results presented in Figure 7. It can be observed that the noise ratio β and regularization factor λ can improve the model performance within a certain range. However, excessive weights for these parameters negatively impact the model by introducing interference. Additionally, we investigated the model dimension d, and the results indicate that insufficient model dimensions fail to adequately capture the characteristics of the data, while overly large dimensions (e.g., $d \geq 512$) do not consistently enhance the model performance.

Model	DB15K				MKG-W				MKG-Y			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
BiQUE	38.34	32.38	41.48	53.23	35.01	29.42	37.01	46.49	36.74	34.82	38.25	42.16
MyGO+Tucker	37.72	30.08	41.26	52.21	36.10	29.78	38.54	47.75	38.44	35.01	39.84	44.19
MyGO+BiQUE	37.43	29.83	41.53	52.05	35.73	29.88	37.38	46.74	37.31	35.23	38.63	42.56
MoMoK+Tucker	39.57	32.38	43.45	54.14	35.89	30.38	37.54	46.13	37.91	35.09	39.20	43.20
MoMoK+BiQUE	40.23	33.43	43.84	54.81	35.23	29.12	36.82	46.12	37.49	35.52	38.86	42.93
M-Hyper (Ours)	**41.25**	**33.64**	**45.01**	**56.09**	**37.02**	**31.24**	**39.16**	**48.84**	**39.46**	**36.02**	**40.92**	**45.22**

Table 4: Baselines with the same decoders, and the embedding dimensions of the BiQUE decoders are kept identical.

Model	Training Time (s)	MRR	Hit@1	Memory Usage (1000MB)	Time Usage (s/epoch)	Params (M)
OTKGE	3505	23.86	18.45	**2.540**	70.1	33.2
MMRNS	7650	32.68	23.01	25.582	25.5	**3.4**
AdaMF	12500	32.51	21.31	10.428	12.5	80.7
IMF	7600	32.25	24.20	3.980	7.6	81.0
MyGO	15900	37.72	30.08	18.128	10.6	23.4
MoMoK	11700	39.57	32.38	5.900	9.8	80.6
M-Hyper	**1200**	**40.75**	**33.14**	2.862	**5.8**	21.5

Table 5: Detailed performance and overhead comparison of M-Hyper at smaller parameter scales.

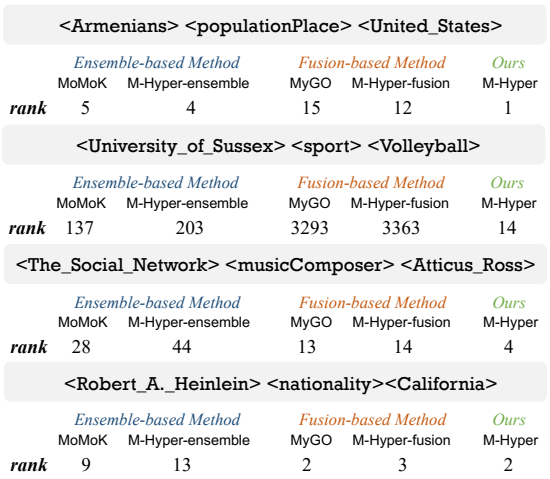

Figure 9: Intuitive cases show the superiority of M-Hyper.

C More Case Analysis Between Paradigms

Specific Relation Performance. To provide a more granular analysis of M-Hyper's advantages, we present the MRR improvements for common relation on DB15K dataset, as shown in Table 3. M-Hyper significantly enhances the performance for 1-to-N relations (e.g., *is_Part_Of, music_Composer*), N-to-1 relations (e.g., *country, language, timeZone*), and N-to-N relations (e.g., *company, associated_Band*). These are challenging for translation-based methods (Xie et al., 2017; Zhao et al., 2022) to address. Additionally, M-Hyper can also achieve at least 6.91% performance improvement in modeling symmetric relationships (e.g., *spouse, different_From*), demonstrating stronger geometric representation capabilities. More case analysis are presented in Appendix C.

M-Hyper-fusion and M-Hyper-ensemble. To further investigate the differences in paradigm shifts, we conduct a more detailed comparison by introducing variations of M-Hyper based on traditional paradigms. Specifically, we keep other modules and the dimension of final embeddings consistent while modifying the score function to create variants: *M-Hyper-fusion* with score function $\phi(h,r,t) = \langle(\mathbf{e}_h^{\mathrm{j}} + \mathbf{r}_r^{\mathrm{T}}) \circledast \mathbf{r}_r^{\mathrm{R}}, \mathbf{e}_t^{\mathrm{j}}\rangle$, and *M-Hyper-ensemble* with score function $\phi(h,r,t) = \sum_m^{|\mathcal{M}|}\langle(\mathbf{e}_h^m + \mathbf{r}_{r,m}^{\mathrm{T}}) \circledast \mathbf{r}_{r,m}^{\mathrm{R}}, \mathbf{e}_t^m\rangle$. Figure 8 illustrates the proportion distribution of different paradigms achieving optimal performance across various relations. It can be observed that M-Hyper achieves the highest proportion in the majority of relationships.

Cases M-Hyper Perform Better. As shown in 9, we present several representative examples of triples under different reasoning requirements. For examples like *(Armenians, populationPlace, United_States)* and *(University_of_Sussex, sport, Volleyball)*, the reasoning results tend to rely more on single-modal features, specifically textual semantic features and analogical reasoning through structural features, respectively. We can find fusion-based methods perform better at preserving the original features, thereby achieving more accurate predictions. In contrast, for relatively long-tail case like *(The_Social_Network, musicComposer, Atticus_Ross)*, and for cases where the answer is relatively sub-optimal like *(Robert_A._Heinlein, nationality, California)*, the model often needs

Model	DB15K					MKG-W				
	Emb. Size	MRR	Hit@1	Hit@3	Hit@10	Emb. Size	MRR	Hit@1	Hit@3	Hit@10
MyGO	800	37.83	30.09	41.31	52.28	800	36.16	29.85	38.53	47.79
MoMoK	800	39.62	32.47	43.44	54.14	800	35.87	30.42	37.58	46.18
M-Hyper	800	**41.21**	**33.68**	**45.06**	**56.14**	800	**37.02**	**31.27**	**39.17**	**48.84**
MyGO	100	36.98	29.14	41.09	51.30	100	35.27	29.10	37.66	46.98
MoMoK	100	38.64	31.97	42.60	53.68	100	35.10	29.38	36.83	45.35
M-Hyper	100	**40.23**	**32.79**	**44.38**	**55.23**	100	**36.21**	**30.45**	**38.47**	**47.98**

Table 6: Performance comparison with baselines at the same embedding sizes as M-Hyper on DB15K and MKG-W datasets.

to collaborate across multiple modalities, such as text and structural information, to infer the answer. Therefore, in this problem type, ensemble-based methods are more suitable for such cooperative reasoning scenarios. At the same time, we observe that M-Hyper surpasses both of these approaches and is more adaptable to diverse and flexible reasoning requirements.

D Comparison under Different Parameter Settings

To demonstrate the effectiveness of M-Hyper and address concerns regarding whether the performance gains are derived from increased parameter count, we present results under three distinct parameter settings:

Baselines with the same decoders. we compare BiQUE (structure-only) against advanced methods equipped with the biquaternion KGE. To ensure fairness, the embedding dimensions for all baselines using the biquaternion KGE decoder are set to be consistent with M-Hyper. Regarding implementation details: for the fusion-based method (MyGO), we split its final fused representation to serve as the input for the biquaternion KGE; for the ensemble-based method (MoMoK), we split the representations of each modality to perform biquaternion KGE calculations separately. As shown in Table 4, M-Hyper consistently outperforms both the unimodal baseline and the BiQUE-enhanced baselines. This confirms that our performance gains stem from the holistic architectural design rather than merely the hypercomplex backbone decoder itself.

Baselines with total training parameters $\geq$ M-Hyper. As shown in Table 5, we provide a detailed comparison of performance and efficiency on the DB15K dataset. It can be observed that compared to most state-of-the-art methods, M-Hyper achieves superior performance even with a significantly smaller number of total training parameters.

Algorithm 1 Noise-powered Self-distillation

Input: Noise rate β; Relation query r; Batch of entity embeddings $\mathcal{E} = \{\mathbf{e}^m\}_{m\in\{s,v,t\}}$.
Output: Distillation Loss $\mathcal{L}_{distill}$

Initialize $\tilde{\mathcal{E}}_{student} \leftarrow \emptyset$
for all $m \in$ {structural, visual, textual} **do**
  Calculate feature mean $\boldsymbol{\varphi}^m$ and variance $\boldsymbol{\mu}^m$
  Generate noise: $\tilde{\mathbf{e}}^m \sim \mathcal{N}(\boldsymbol{\varphi}^m, \boldsymbol{\mu}^m)$
  Sample binary mask $\mathbf{M}$ with probability β
  Inject noise: $\mathbf{e}^{m\prime} \leftarrow \mathbf{e}^m + \mathbf{M} \odot \tilde{\mathbf{e}}^m$
  $\tilde{\mathcal{E}}_{student} \leftarrow \tilde{\mathcal{E}}_{student} \cup \{\mathbf{e}^{m\prime}\}$
end for
Compute r-aware weights w^m via Eq. (6)
Fuse clean embeddings: $\hat{\mathbf{e}}^{\mathrm{j}} \leftarrow \sum w^m \mathbf{e}^m$
Compute weights and fuse noisy embeddings $\tilde{\mathcal{E}}_{student}$
Obtain student embedding: $\hat{\mathbf{e}}^{\mathrm{j}\prime} \leftarrow \mathrm{R2MF}(\tilde{\mathcal{E}}_{student}, r)$
Calculate MSE loss: $\mathcal{L}_{distill} = \|\hat{\mathbf{e}}^{\mathrm{j}} - \hat{\mathbf{e}}^{\mathrm{j}\prime}\|^2$
return $\mathcal{L}_{distill}$

Baselines with embedding dimensions equal to M-Hyper. To ensure a fair comparison, for methods utilizing hypercomplex decoders, we aligned the total dimension of their hypercomplex components with the corresponding real-valued models. We present the comparative results under two embedding size settings (800 and 100) in Table 6. The experimental results demonstrate that M-Hyper consistently outperforms other baselines under identical embedding dimensions. This indicates that the performance improvement is not solely attributed to an increase in parameter count.

E Pseudocode of "Noise-powered Self-distillation"

As shown in pseudocode 1, we have provided a detailed description of the process of the "Noise-powered Self-distillation" module.

Dataset	$\|\mathcal{E}\|$	$\|\mathcal{R}\|$	#Train	#Valid	#Test	image		Text	
						Num	Dim	Num	Dim
DB15K	12842	279	79222	9902	9904	12818	4096	9078	768
MKG-W	15000	169	34196	4276	4274	14463	383	14123	384
MKG-Y	15000	28	21310	2665	2663	14244	383	12305	384

Table 7: The statistics of three MMKG benchmarks.

F Dataset Statistics

The statistical details of dataset are shown in Table 7.