MIDF: 基于模态交互和关系引导决策融合的多模态知识图谱补全

曾啸尘，赵晖[†]，英迪
新疆大学，计算机科学与技术学院/ 新疆，乌鲁木齐
107552304197@stu.xju.edu.cn，277875592@qq.com，yd77889977@163.com

摘要

多模态知识图补全（MMKGC）通过融合实体间的结构化语义信息与多模态特征，从给定的多模态知识图谱（MMKG）中发现未观察到的潜在事实。然而，现有方法普遍忽略了实体表示过程中不同模态的交互，同时缺乏对补全过程中模态之间互补性的关注。为了解决这些不足，我们提出了一种新的模型MIDF（模态交互和决策融合）来处理多模态的交互和互补。该模型首先设计了一个实体多模态交互融合模块，将实体的图像和文本特征提前交互后，再与结构特征进行融合，充分学习实体的嵌入。为了在补全过程中进一步利用不同模态之间的互补性，我们设计了关系引导的决策融合模块。通过使用不同模态的预测结果以及关系引导的权重，进一步利用模态的互补性，融合预测结果。在DB15K和MKG-W上的广泛实验证明，我们的MIDF优于现有的最先进的模型，证明了我们方法的有效性。

关键词： 多模态知识图谱 ； 知识图谱补全 ； 多模态融合

MIDF: Multimodal Knowledge Graph Completion Based on Modal Interaction and relationship guided decision fusion

Xiaoceng Zeng，Hui Zhao* ，Di Ying
Xinjiang University, School of Computer Science and Technology/Urumqi, Xinjiang
107552304197@stu.xju.edu.cn，277875592@qq.com，yd77889977@163.com

Abstract

Multi-modal Knowledge Graph Completion (MMKGC) discovers unobserved potential facts from a given Multi-modal Knowledge Graph (MMKG) by integrating structured semantic information and multi-modal features among entities. However, existing methods generally neglect the interaction between different modalities during entity representation and lack attention to the complementarity among modalities during the completion process. To address these shortcomings, we propose a new model, MIDF (Modal Interaction and Decision Fusion), to handle multi-modal interaction and complementarity. The model first designs an entity multi-modal interaction fusion module, which interacts the image and text features of entities in advance and then fuses them with structural features to fully learn the embeddings of entities. To further utilize the complementarity among different modalities during the completion process, we design a relationship-guided decision fusion module. By using the prediction results of different

[†] 通讯作者
基金项目：新疆维吾尔自治区重点研发计划项目（2023B01032）

modalities and relationship-guided weights, we further leverage the complementarity of modalities and fuse the prediction results. Extensive experiments on DB15K and MKG-W demonstrate that our MIDF outperforms the existing state-of-the-art models, proving the effectiveness of our method.

Keywords: Multi-modal Knowledge Graph , Knowledge Graph Completion , Multimodal fusion

1 引言

多模态知识图谱(Multimodal Knowledge Graphs, MMKGs)(Liu et al., 2019)作为传统知识图谱的扩展形态，通过将文本、图像等多模态特征与结构化三元组(头实体，关系，尾实体)进行语义关联，构建了实体间多维度特征空间。这种融合表征范式为推荐系统(Sun et al., 2020)、视觉问答(Hudson et al., 2019)以及大语言模型(Chen et al., 2023; Dong et al., 2024)等下游任务提供了更可靠的知识支撑。然而，由于多模态的语料库难以搜集，现有的多模态知识图谱通常面临着不完整问题。因此，知识图谱补全(Knowledge Graph Completion, KGC)(Bordes et al., 2013; Sun et al., 2019)技术通过挖掘图谱潜在关联发现潜在事实，已成为知识图谱的重要研究方向。

传统的知识图补全方法(Cao et al., 2022; Li et al., 2023; Xie et al., 2016; Mousselly et al., 2018)通常学习结构嵌入来建模三元组并评估三元组的有效性。然而，对于多模态知识图，仅依靠结构嵌入不足以捕捉模态之间的复杂关系和相互作用。有必要有效地整合来自不同模态的信息，以实现对缺失实体的更准确的预测。在已有的多模态知识图补全研究中，通常采用实体的多模态信息作为编码实体表示的补充数据，利用预训练模型从不同模态中提取特征。然后，将编码的实体表示与关系表示连接起来，输入到三元组编码器中进行三重预测。然而，这些方法仍然面临实体模态交互和模态互补问题。

模态交互：如图1所示，现实中多模态知识图谱中实体的不同模态之间存在联系。现有的多模态知识图谱补全方法通常使用点积、拼接等简单直接的方式进行实体多模态特征融合，获取实体嵌入，忽略了实体不同模态之间存在复杂的交互。一些方法直接使用预训练模型获取的原始模态信息，再通过Transformer进行实体多模态融合。然而，由于直接输入所有模态信息，这些方法通常会丢失实体的细节信息，并受到噪声影响。**模态互补**：在多模态知识图谱补全过程中，不同模态预测结果不同。以图1为例，在预测Lionel Messi所在的俱乐部时，文本模态可能在FC Barcelona和Argentina national team之间选择。而视觉模态难以发现细粒度的信息，只能判断Lionel Messi是个足球运动员。只有通过整合这些知识，才能得出正确答案。一些研究(Zhao et al., 2022)尝试使用决策融合的方法，试图解决此问题。这些方法分别提取不同模态的特征，接着使用不同模态得分函数对三元组得分进行预测。然而，这些方法在决策过程中使用简单的加权计算最终得分，没有考虑预测时模态在不同关系下不同的重要性。

Figure 1: 实体“Lionel Messi”的文本和视觉模态间具有联系。

为了获取模态交互增强的实体嵌入，学习模态之间的关联，以及在预测时合理利用不同模态之间的互补性，我们提出了一种新的多模态知识图谱补全模型MIDF。MIDF首先设计了一个文本-视觉提前交互层，将从预训练模型中提取的实体的文本、视觉表示输入到提前交互层中进行交互。通过图像-文本的早期交互，实现文本、图像跨模态语义增强，解决单一模态的模糊性。例如文本描述中的"FC Barcelona"与图像中球衣上的徽章区域建立细粒度语义关联。为了进一步融合实体结构信息，我们再将交互后的文本、视觉表示和预训练提取的结构表示拼接起来，输入多模态融合层，获得融合后的实体多模态特征和结构、文本、视觉特征。针对预测时不同模态的互补性，我们通过关系引导加权每个模态的预测结果得到最终预测。具体来说，我们使用实体的多模态特征和模态交互后的结构、文本、视觉特征作为输入得到每个模态的预测结果，再通过关系引导计算出模态权重。总的来说，我们的贡献在于四个方面:

1.我们提出了一种新的多模态知识图谱补全模型MIDF，该模型通过实体多模态信息交互，获得增强的模态语义，解决实体单一模态的模糊性。并且在预测阶段，通过关系引导决策融合，进一步利用不同模态之间的互补性。

2.提出了一个实体多模态交互融合模块，将实体的文本-图像模态进行初步的交互，再和结构信息进行融合，捕捉实体模态交互，获得实体有效表示。

3.设计了关系引导的决策融合模块，利用关系引导决策阶段融合，合理利用模态互补性提升决策质量。

4.我们在两个公开数据集上进行了广泛的实验，结果表明我们的模型明显优于现有的方法。

2 相关工作

2.1 知识图谱补全

传统的知识图谱补全（KGC）任务旨在通过已知知识图谱（KG）发现未观察到的三元组。现有的方法通常都是基于嵌入的，通过将KG中的实体和关系嵌入到低维向量空间，并根据现有的三元组结构学习嵌入。再设计不同的评分函数以评测三元组的合理性，以为正三元组分配更高的分数，并为负三元组分配较低的分数为目标。现有的知识图谱补全（KGC）可以分为两大类，基于翻译距离的方法和基于语义匹配的方法。基于翻译距离的方法将三元组结构建模为从头部实体到尾部实体的关系翻译，设计了基于距离的评分函数作为合理性测量。如TransE(Bordes et al., 2013)通过一个简单的函数，即头实体、关系和尾实体之间的向量关系应该满足h+r≈t。此外，RotatE(Sun et al., 2019)、OTE(Tang et al., 2019)和PairRE(Chao et al., 2020)在此基础上进一步优化了评分函数。基于语义匹配的方法通常通过计算头实体、关系和尾实体之间的相似度作为合理性度量。如DistMult(Yang et al., 2014)、ComplEx(Trouillon et al., 2016)、TuckER(Balažević et al., 2019)使用基于张量分解的方法作为可信性度量，还有一些方法通过深度神经网络提取结构语义。

2.2 多模态知识图谱补全

多模态知识图谱(MMKGs)通过整合结构、文本、图像、音频等多模态信息，增强了知识表示的丰富性和推理能力。多模态知识图谱补全通常在单一模态的方法上进行改进，使用从预训练模型中获得的多个模态特征，其目标是通过协同利用多模态信息预测缺失的三元组。现有方法主要从以下三个方面改进:(1)多模态融合，(2)综合决策，(3)负采样。第一种方法(Cao et al., 2022; Chen et al., 2024; Lee et al., 2023; Wang et al., 2021; Wang et al., 2019; Xie et al., 2016)认为多模态信息的有效融合是MMKGC的核心，现有方法通过一些复杂的机制实现。比如AdaMF-MAT(Zhang et al., 2024)框架通过自适应模态权重动态调整不同模态的贡献，并引入对抗训练生成合成样本以缓解模态不平衡问题。第二种方法(Zhao et al., 2022; Li et al., 2023)在决策阶段整合多模态预测结果，输出联合决策。例如，MoSE(Zhao et al., 2022)利用结构、文本和视觉数据来训练三个KGC 模型，并使用集成策略进行联合预测。负采样是KGC模型训练的关键环节，第三种方法(Xu et al., 2022; Zhang et al., 2023; Zhang et al., 2022)针对MMKGC中多模态特性提出改进。比如KBGan(Cai et al., 2017)使用对抗性方法来增强MMKGC 模型，将对抗训练引入MMKGC，向模态嵌入添加噪声。

3 任务定义

一个多模态知识图(MMKG)可以表示为G={E,R,T,M}，其中E,R是实体集和关系集。T={(h,r,t)|h,t∈E,r∈R}是三元组集，表明实体h 通过关系r与实体t相关。M表示实体的多个模态，为每个实体e∈E提供图像、文本等多模态描述集合。知识图补全(KGC)的主要目的是学习一个得分函数F(h,r,t) :E×R×E→ℝ，它通过标量分数来衡量三元组(h,r,t) 的合理性。在KGC 模型中，实体和关系对应于嵌入，三元组分数在这些嵌入上定义，更倾向于正三元组的高分和负三元组的较低分数，实现正负样本对比学习—即最大化正三元组得分同时最小化负采样三元组得分。扩展到MMKG，多模态知识图补全(MMKGC)将进一步考虑每个实体e的多模态信息M (e)，现有方法通常采用多模态融合策略：首先为各模态生成独立嵌入{e_m,m∈M}，再通过拼接、加权平均等方式融合为统一表示，以增强它们的嵌入。当前研究聚焦于设计高效的多模态融合机制，通过引入注意力机制、图神经网络等先进技术，实现模态间互补信息的自适应整合，从而显著提升三元组预测的准确性。在推理阶段，具有预测给定查询的缺失实体的MMKGC 模型(?,r,t)或(h,r,?)。比如尾实体预测(h,r,?)，MMKGC 模型将每个实体e∈E视为候选实体并计算其对应的分数(h,r,e)。此外，该模型通过黄金答案(h,r,t)与所有候选者的排名来评估，这意味着将基于排名的指标将用于性能评估。评估以平均倒数排名(MRR)和Hits@k(k=1,3,10)为核心指标，同时计算头实体预测(?,r,t)与尾实体预测(h,r,?)的综合性能。

4 方法

在本节中，我们将详细介绍我们所提出的MIDF模型。如图2所示，我们的模型分两个阶段，旨在实现实体跨模态语义增强和提升决策阶段的准确性。首先先将实体的文本和视觉嵌入通过模态交互层获得交互后的文本和视觉特征，接着将结构和交互后的文本和视觉特征输入模态融合模块获得实体多模态、以及其他模态特征。在获得了实体的4种不同特征之后，在决策阶段，通过评分函数获得各个模态的得分，最后在关系的引导下进行决策融合。

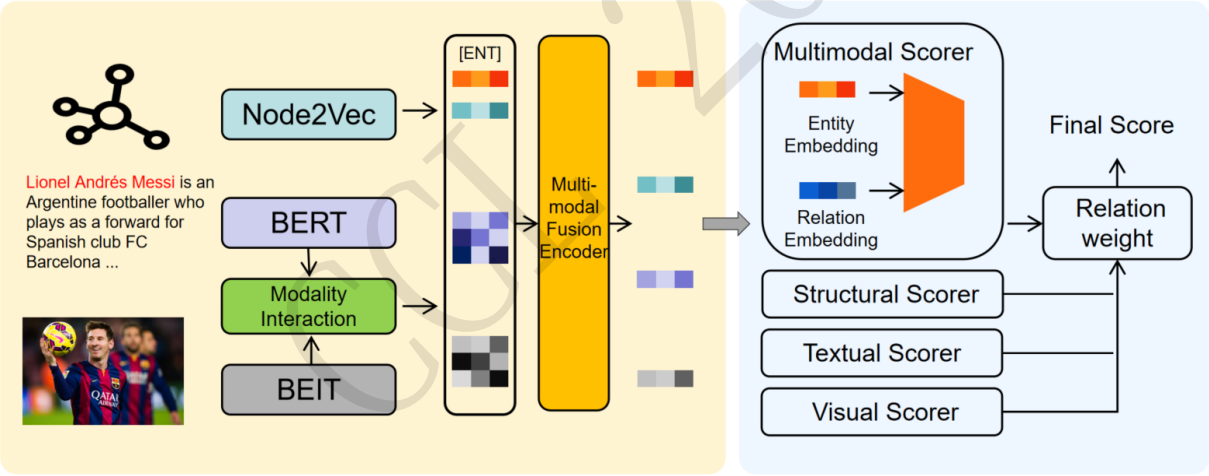

Figure 2: 基于模态交互和决策融合MIDF的整体框架，主要由模态交互层，多模态融合层，关系引导的决策融合模块组成。

4.1 模态交互与融合

在模态融合前，我们先通过不同预训练模型来获得实体的单模态表示。首先我们使用预训练的BERT(Devlin et al., 2019)对实体的文本描述进行编码以获取文本模态的文本标记。具体来说，我们先删除文本描述中的停用词，在根据文本词汇表生成对应的tokens。并且对于标记中可能重复的tokens的问题，类似于MyGO(Zhang et al., 2024)，我们先统计tokens出现的频率，保留文本中最常见的x个tokens，接着我们计算每个token对应的嵌入d。对于视觉模态来说，我们使用预训练的BEIT(Peng et al., 2022)对图像进行编码。同样的，先基于码本生成视

觉tokens，再保留图像中最常见的y个tokens。对于实体的结构信息，不同于之前的方法将其视为可学习的张量，我们的方法和C2RS(Shu et al., 2025)类似。先将整个知识图谱视为一个没有属性的有向图，这样只要关注实体和实体间关系的结构。然后使用Node2vec(Grover et al., 2016)通过随机游走策略提取每个实体的结构信息，在实体编码时视为结构tokens。前面提到的所有预训练模型都可以被替换，以获得更好的嵌入。同时，在训练期间，所有预训练模型都被冻结不更新参数。在获得了实体的单模态实体后，我们先将这些表示通过线性投影层，映射到统一的向量空间：

$$s = \mathrm{P}_s(\mathbf{e}_s^0), \quad t = \mathrm{P}_t(\mathbf{e}_t^0), \quad v = \mathrm{P}_v(\mathbf{e}_v^0) \tag{1}$$

其中s,t,v分别表示实体的不同模态，包括结构特征、文本特征、视觉特征，$\mathbf{e}_m^0$表示从预训练模型中获得的初始特征，P_m表示对应模态的线性投影层。多模态实体的不同模态之间存在联系，例如文本描述中的斑马的条纹和图像中的斑马的纹理区域是相互关联的。在实体多模态融合中，学习到这种相关性是非常重要的，它可以使实体的文本描述和图像区域建立联系并且抑制无关特征，如文本中的冗余修饰词或图像的背景干扰。然而现有的方法通常直接融合初始特征，忽视了这种不同模态的内在相关性，影响了实体嵌入的获取。为此我们设计了一个模态提前交互层MI，通过注意力机制使文本和图像在浅层即建立关系。具体来说，我们获得初始的图像和文本特征之后，将两者拼接起来并通过transformer(Vaswani et al., 2017)中的注意力机制捕捉模态间的长距离依赖。这种方法能保留更多原始信息中的细节（如图像中的物体关系、文本中的语义逻辑），使融合后的实体特征更细粒度地表达模态间关联。我们先将文本特征和视觉特征拼接作为输入：

$$M_{\text{input}} = (t_1, t_2, t_3, \ldots, t_n, v_1, v_2, v_3, \ldots, v_n) \tag{2}$$

其中t_i表示t中第i个文本token，v_i表示v中第i个视觉token。该序列输入到模态交互层中，该模态交互层由一层的transformer编码器构成。通过transformer中的注意力机制，我们能使文本和视觉模态在融合前建立细粒度语义联系。最后得到输出：

$$M_{\text{output}} = (t_1', t_2', t_3', \ldots, t_n', v_1', v_2', v_3', \ldots, v_n') \tag{3}$$

其中t_i'，v_i'表示交互后的第i个文本token和视觉token。得到交互后的文本和视觉tokens后，为了进一步注入增强知识并获得实体多模态特征，我们再进一步融合,设计了一个多模态融合层。多模态融合层通过transformer来捕获实体的多模态表示，并且进一步提升图像、文本特征的表征能力，提升后续的决策能力。具体来说，我们先将多个模态tokens拼接为一个序列：

$$X_{\text{input}} = \left([\text{ENT}], s, t_1', t_2', \ldots, t_n', v_1', v_2', \ldots, v_n'\right) \tag{4}$$

其中[ENT]是个特殊标记，用来捕获实体的多模态融合特征，在训练阶段是可学习的嵌入，s表示前面提到的实体的结构token。我们将拼接的序列输入多模态融合层，获得包括实体多模态特征和不同单模态特征的输出：

$$X_{\text{output}} = (e_m, e_s, e_{t_1}, e_{t_2}, \ldots, e_{t_n}, e_{v_1}, e_{v_2}, \ldots, e_{v_n}) \tag{5}$$

其中e_m表示融合了其他模态的实体多模态嵌入，用于后续的处理。对于三个模态，我们根据结构、图像和文本标记的位置拆分多模态实体编码器的输出，然后对这些标记分别执行平均池化以获得它们的向量表示e_s、e_t和e_v。

4.2 关系引导的决策融合

基于上述设计，我们获得了实体交互融合后的多模态特征、结构、文本和视觉特征。为了进一步降低模态冲突，提升预测的准确性，我们尝试利用不同模态的互补性。现有的多模态方法主要侧重于将不同的模态表示投影到一个统一的空间中，并预测模态之间的共性，这将无法保留模态特定的知识。在多模态知识图谱补全中，关系通常影响模态间信息的分布。例如在“颜色”的关系语境中，实体间的视觉模态通常更相关，该模态的预测结果也更准确。基于这一观察，我们在决策融合阶段通过使用交互后的模态特征和利用关系引导决策融合FD。具体来说，

我们先获得实体各个模态嵌入和关系嵌入，使用评分函数S（h，r，t）通过生成标量分数来衡量三元组的合理性。在MIDF中我们使用Tucker(Balažević et al., 2019)作为我们的评分函数，以多模态嵌入为例表示为：

$$\mathrm{S}(\mathrm{h_m}, \mathrm{r}, \mathrm{t_m}) = W \times_1 \mathrm{h_m} \times_2 \mathrm{r} \times_3 \mathrm{t_m} \tag{6}$$

其中$\times_i$表示沿第i个模式的张量积，W是训练过程中学习到的核心张量，h_m和t_m表示头实体和尾实体的多模态嵌入。我们用每个三元组的交叉熵损失来训练我们的模型。在屏蔽尾实体时，我们将t_m视为黄金标签，头部预测与其相同。根据分数对所有实体进行排名，因此，训练目标是交叉熵损失：

$$L^m_{\text{head}} = -\sum_{(h,r,t)\in T} \log \frac{\exp\left(S(h_m, r, t_m)\right)}{\sum_{t'\in\mathcal{E}} \exp\left(S(h_m, r, t_m)\right)} \tag{7}$$

$$L^m_{\text{tail}} = -\sum_{(h,r,t)\in T} \log \frac{\exp\left(S(h_m, r, t_m)\right)}{\sum_{h'\in\mathcal{E}} \exp\left(S(h_m, r, t_m)\right)} \tag{8}$$

$$L_m = L^m_{\text{head}} + L^m_{\text{tail}} \tag{9}$$

其中L_m表示多模态m的交叉熵损失。同样的，我们使用Tucker作为我们的评分函数，获得结构、文本、视觉模态的损失L_s、L_t、L_v。传统的决策融合使用简单的自适应权重加权各个模态的损失，作为联合损失。这种方法忽视了不同关系下，实体的不同模态在知识图谱补全任务中的作用不同。因此，我们使用前面的设计得到的关系嵌入和实体各个模态特征，通过注意力机制计算实体不同模态在不同关系下的权重。给定一个缺失三元组(e，r，?)，实体e的权重：

$$a_i = \frac{\exp\left(\frac{\sum(e_i \odot r)}{\sqrt{d}}\right)}{\sum_{j\in K} \exp\left(\frac{\sum(e_j \odot r)}{\sqrt{d}}\right)} \tag{10}$$

其中e_i表示实体的模态信息，包括e_m、e_s、e_t和e_v，K表示四个不同模态。r表示关系，$\odot$表示Hadamard乘积。a_i表示对应模态在关系r中的重要性，我们将其作为不同模态预测结果的权重。通过关系引导的各个模态的加权，由此我们得到了整个模型的联合损失：

$$L_{\text{joint}} = a_m L_m + a_s L_s + a_t L_t + a_v L_v \tag{11}$$

5 实验

在本节中，我们将进行综合实验来评估MIDF的性能。我们首先详细介绍我们的实验设置，然后进行全面的结果分析。

5.1 实验设置

5.1.1 数据集

在本文中，我们使用两个公共MMKGC基准DB15K(Liu et al., 2019)和MKG-W(Xu et al., 2022)来评估模型性能。DB15K源自DBpedia(Lehmann et al., 2015)，MKG-W是Wikidata(Vrandečić et al., 2014)的一个子集。两者都由图像和文本描述组成，具有丰富的多模态上下文。所有数据集都由三种模式组成：结构三元组、实体图像和实体描述，我们利用关系三元组作为结构特征，实体图像作为视觉特征，并从数据中提取实体描述作为文本特征。数据集的详细信息见表1，每种模式的原始数据都是从其官方发布来源获得的。在训练过程中，每个数据集被划分为训练集、验证集和测试集，分割比为8:1:1。

5.1.2 评估标准

我们对数据集进行链接预测任务，这是MMKGC中的主要任务。在现有工作的基础上，我们使用基于秩的度量，如平均倒数秩（MRR）和Hit@K（K=1,3,10）以评估结果。此外，我们

Dataset	$\|E\|$	$\|R\|$	Train	Valid	Test	$\|T\|$	$\|V\|$
DB15K	12842	279	79222	9902	9904	12818	12842
MKG-W	15000	169	34196	4276	4274	14463	14123

Table 1: 数据集的统计信息。

在预测结果中使用滤波器设置来去除训练数据中存在的候选三元组，以进行公平比较。MRR和Hits@K 可以表示如下：

$$\text{MRR} = \frac{1}{|T_{\text{test}}|} \sum_{i=1}^{|T_{\text{test}}|} \left(\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}} \right) \tag{12}$$

$$\text{Hits@}K = \frac{1}{|T_{\text{test}}|} \sum_{i=1}^{|T_{\text{test}}|} \left(\mathbf{1}(r_{h,i} \leq K) + \mathbf{1}(r_{t,i} \leq K) \right) \tag{13}$$

其中$r_{h,i}$和$r_{t,i}$是预测的头部实体排名和预测的尾部实体排名，T_{test}是测试三元组。

5.1.3 基线方法

在我们的实验中，我们与22种不同的最先进的基线进行全面的比较和分析。从模态的角度出发，基线可以被分为两类。对于单模态，我们只使用三元组的结构信息进行训练，对于多模态方法，我们添加文本信息和视觉信息来训练模型。

(1)单模态KGC方法:TransE(Bordes et al., 2013)、DistMult(Yang et al., 2014)、ComplEx(Trouillon et al., 2016)、RotatE(Sun et al., 2019)、PairRE(Chao et al., 2020)、GC-OTE(Tang et al., 2019)和Tucker(Balažević et al., 2019)。这些传统方法在模型设计中使用不同的评分函数并只考虑结构信息，例如DistMult通过对角张量乘法捕捉实体和关系之间的相互作用，RotatE引入实体之间的旋转操作来表示复空间中的关系。

(2)多模态方法:IKRL(Xie et al., 2016)、TBKGC(Mousselly et al., 2018)、TransAE(Wang et al., 2019)、MMKRL(Lu et al., 2022)、RSME(Wang et al., 2021)、VBKGC(Zhang et al., 2022)、OTKGE(Cao et al., 2022)、IMF(Li et al., 2023)、QEB(Wang et al., 2023)、VISTA(Lee et al., 2023)、AdaMF(Zhang et al., 2024)、MyGO(Zhang et al., 2024)、NativE(Zhang et al., 2024)和C2RS(Shu et al., 2025)。这些方法考虑了MMKGC模型中的图像和文本信息，并采用了不同的的多模态融合方法。TransAE使用TransE评分功能在融合图像和文本模态之前分别嵌入它们，IMF学习了一种具有四种不同模式的多模态知识图完成模型，以实现联合决策。

5.1.4 实施细节

我们用PyTorch实现了MIDF，并将训练周期数设置为1500，批处理大小设置为4096。使用BEIT和BERT的标记器作为视觉/文本标记器。对于DB15K，嵌入维度设置为256，每个实体的选定文本标记数量设置为8。对于MKG-W，嵌入维度设置为128，每个实体的选定文本标记数量设置为24。对于所有数据集，实体的选定视觉标记的数量设置为8。对于Bert，嵌入维度为32，对于Beit，嵌入维度为768，进过投影后，统一投影至256。对于Node2Vec，嵌入维度设置为256。在设置的随机游走策略中，每个节点执行20次随机游走，每次游走长度为5。在训练过程中，Node2Vec最多考虑10个相邻节点作为当前节点的上下文，最小词频设置为1。对于模态交互层和多模态融合层，我们都采用1个具有8个注意头的transformer层用于多模态实体编码器，同时2个具有8个子注意头的transformer层用于三重编码器。我们使用Adam优化器优化模型，学习率为5e-4，dropout概率为0.4。所有实验在一个具有Ubuntu20.04.1操作系统和一张NVIDA RTX3090TI GPU的Linux服务器上进行。

5.2 主要结果

MIDF和基线方法在链接预测的主要结果如表2所示。我们可以从中分析出一些结果。首先我们可以看到，MIDF在四个指标上优于所有的基线模型，在两个数据集上都取得了最优的结

果。这说明了我们提出的在多模态融合前进行模态提前交互，有效地提升了实体表示的质量。并且水平比较每个指标的提升可以发现，MIDF对Hit@1和MRR的改进明显高于Hit@10 和其他指标。这说明了我们设计的关系引导的决策融合在预测准确性上有明显提升。其次，与其他基于复杂的多模态融合方法或使用决策融合的多模态知识图谱补全模型相比，MIDF的模块虽然简单，但是表现得更优秀，这说明了我们设计的模型的有效和高效。与其他模型相比，我们的模型通过使用充分交互融合后的不同特征进行决策融合，有效利用了不同模态之间的相关性和互补性。第三，MIDF在DB15K 上的性能提升大于MKG-Y 的性能提升。这是因为与MKG-Y相比，DB15K中实体所拥有的图片和文本信息更充分，这有助于模型充分利用多模态信息。

Category	Model	DB15K				MKG-W			
		MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Uni-modal	TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23
	DistMult	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86
	ComplEx	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73
	RotatE	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73
	PairRE	31.13	21.62	35.91	49.30	34.40	28.24	36.71	46.04
	GC-OTE	31.85	22.11	36.52	51.18	33.92	26.55	35.96	46.05
	Tucker	33.86	25.33	37.91	50.38	30.39	24.44	32.91	41.25
Multi-modal	IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07
	TBKGC	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24
	TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72
	MMKRL	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69
	RSME	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43
	VBKGC	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88
	OTKGE	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88
	IMF	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44
	QEB	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32
	VISTA	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61
	AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.21
	MANS	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78
	MMRNS	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47
	NativE	37.16	28.01	42.25	54.13	36.58	29.56	38.57	47.81
	MyGO	37.72	30.08	41.26	52.21	36.10	29.78	38.54	47.75
	C2RS	39.65	31.93	43.16	54.57	39.75	32.85	42.37	52.84
MIDF		**41.10**	**33.27**	**44.92**	**55.94**	**40.58**	**33.83**	**43.02**	**52.85**

Table 2: DB15K和MKG-W的MIDF和主要MMKGC结果。我们在表中列出了每种方法的类型（单模态/多模态）。最好的结果用粗体标记。

5.3 消融实验

Setting	MRR	Hit@1	Hit@3	Hit@10
w/o image	40.51	32.60	44.23	55.69
w/o text	40.73	32.69	44.61	56.14
w/o MI	39.98	32.30	43.56	54.56
w/o FD	40.11	32.36	43.66	55.05
Full MIDF	41.10	33.27	44.92	55.94

Table 3: DB15K的消融研究结果。我们进行了三组实验，分别验证不同模态、模态交互模块和决策融合模块的有效性。

为了证明我们设计的模块在MIDF中的有效性，我们设计了一系列的消融实验。我们分别从模态信息和模块设计两个角度，在不同设置中去除了一些模块并执行MMKGC 实验。实验结果如表3所示。在消融实验中，我们进行了三组实验，通过移除相应的关键组件来验证MIDF中每个模块的贡献。从模态信息的移除结果来看，我们可以得出结论，每种模态都对预测结果有多提升。尽管相比较文本模态，图像模态的作用较小。我们通过移除模态交互层证明了其作用，因为它的移除导致了模型的性能下降。而在关于决策融合模块的实验证明了，我们的关系加权带来了巨大的性能提升，尤其是在准确性方面。

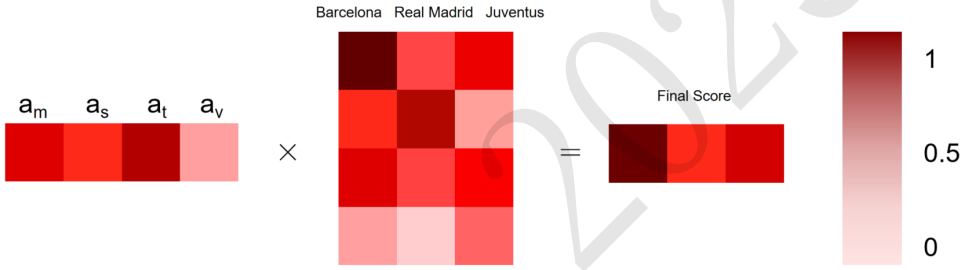

Figure 3: 决策融合中的预测分数计算。

数据集	模型	文本到图像检索			图像到文本检索		
		R@1	R@5	R@10	R@1	R@5	R@10
DB15K	MIDF(w/o MI)	51.5	79.6	88.3	31.6	58.4	72.3
	MIDF	95.3	97.2	99.8	83.1	96.0	98.2

Table 4: 图文检索结果。

5.4 可视化

为了进一步说明我们的关系引导的决策融合的有效性，我们选择了（Lionel Messi，team）的情况，并将每个模态的预测分数可视化，如图3所示,其中颜色深浅表示数字大小。很明显在多模态中，只依靠结构信息进行决策的是不可避免出现错误的。这证明了关系在决策融合中的有效性，我们提出的方法对捕捉不同模态之间的互补是有益的。

为了评估我们的MIDF模型是否有效地解决了实体模态交互的问题，我们将图像文本检索任务作为评估方法。检索结果的准确性可以反映模型模态交互的有效性。通过量化模型对查询模态与目标模态在语义上匹配的能力，我们可以间接评估模型是否成功地捕获了不同模态之间的语义关联。具体来说，我们进行了两个任务，即图像到文本检索和文本到图像检索。模型性能使用R@k进行评估。结果如表4所示，经过模态交互后，实体的图像文本模态产生了关联。

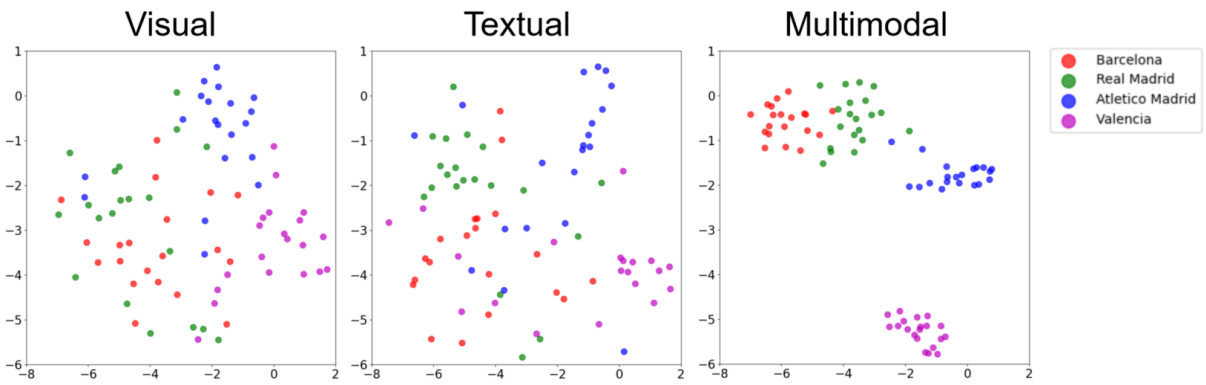

Figure 4: 每个节点代表一个足球运动员，不同的颜色代表球员所在俱乐部。

为了直观说明我们设计的模态提前交互是否有效地使实体的文本-视觉模态交互，我们应用t-SNE进行降维和可视化不同球队的球员的模态交互后的上下文实体表示，以及融合后的实体多模态表示。如图4所示，我们可以发现经过模态交互后的表示明显更容易区分。这表明我们的模型可以区分这些相似的实体。

6 结论

在本文中，我们提出了一个基于多模态交互和决策融合的多模态知识图谱补全方法。我们从优化多模态特征交互和利用不同模态的互补性两个角度出发，实现有效的多模态知识图谱补全。具体来说，我们首先设计了一个多模态交互融合模块，通过使除结构信息以外的模态进行初步的交互，再和结构信息拼接通过多模态融合层进行融合，让各个模态得到充分交互。然后使用融合后的多模态特征、结构、文本、视觉特征实现决策融合。在两个基准数据集上，我们进行了广泛的实验，实验结果证明了我们的模型的有效性。对于未来的工作，我们可能会更重视结构信息和三元组中的关系，进一步提升多模态特征的语义的准确性。或者进一步优化模型，适当的添加一些模块以提升模型性能，优化模块以减少模型参数量等。

参考文献

Balažević I, Allen C, Hospedales T M. 2019. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*.

Bordes, Antoine and Usunier, Nicolas and Garcia-Duran, Alberto and Weston, Jason and Yakhnenko, Oksana. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*,26.

Cai, Liwei and Wang, William Yang. 2017. Kbgan: Adversarial learning for knowledge graph embeddings. *arXiv preprint arXiv:1711.04071*.

Cao, Zongsheng and Xu, Qianqian and Yang, Zhiyong and He, Yuan and Cao, Xiaochun and Huang, Qingming. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in neural information processing systems*,35: 39090-39102.

Chao, Linlin and He, Jianshan and Wang, Taifeng and Chu, Wei. 2020. Pairre: Knowledge graph embeddings via paired relation vectors. *arXiv preprint arXiv:2011.03798*.

Chen, Zhuo and Fang, Yin and Zhang, Yichi and Guo, Lingbing and Chen, Jiaoyan and Pan, Jeff Z and Chen, Huajun and Zhang, Wen. 2024. Noise-powered Multi-modal Knowledge Graph Representation Framework. *arXiv preprint arXiv:2403.06832*.

Chen, Zhuo and Zhang, Wen and Huang, Yufeng and Chen, Mingyang and Geng, Yuxia and Yu, Hongtao and Bi, Zhen and Zhang, Yichi and Yao, Zhen and Song, Wenting and Wu, Xinliang and Yang, Yi and Chen, Mingyi and Lian, Zhaoyang and Li, Yingying and Cheng, Lei and Chen, Huajun. 2023. Tele-Knowledge Pre-training for Fault Analysis. *2023 IEEE 39th International Conference on Data Engineering (ICDE)*,pages 3453-3466.

Chen, Zhuo and Zhang, Yichi and Fang, Yin and Geng, Yuxia and Guo, Lingbing and Chen, Xiang and Li, Qian and Zhang, Wen and Chen, Jiaoyan and Zhu, Yushan and others. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*,pages:4171–4186.

Dong, Junnan and Zhang, Qinggang and Zhou, Huachi and Zha, Daochen and Zheng, Pai and Huang, Xiao. 2024. Modality-aware integration with large language models for knowledge-based visual question answering. *arXiv preprint arXiv:2402.12728*.

Esser, Patrick and Rombach, Robin and Ommer, Bjorn. 2021. Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,pages:12873-12883.

Goodfellow, Ian J and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua. 2014. Generative adversarial nets. *Advances in neural information processing systems*,27.

Grover, Aditya and Leskovec, Jure. 2016. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*,pages:855-864.

Hudson, Drew A and Manning, Christopher D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,pages:6700–6709.

Je, Sang-Hyun. 2022. Entity aware negative sampling with auxiliary loss of false negative prediction for knowledge graph embedding. *arXiv preprint arXiv:2210.06242*.

Kingma, Diederik P and Ba, Jimmy. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, Jaejun and Chung, Chanyoung and Lee, Hochang and Jo, Sungho and Whang, Joyce. 2023. VISTA: Visual-textual knowledge graph representation learning. *Findings of the association for computational linguistics: EMNLP 2023*,pages:7314-7328.

Lehmann, Jens and Isele, Robert and Jakob, Max and Jentzsch, Anja and Kontokostas, Dimitris and Mendes, Pablo N and Hellmann, Sebastian and Morsey, Mohamed and Van Kleef, Patrick and Auer, Sören and others. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*,6(2): 167-195,SAGE Publications Sage UK: London, England.

Li, Xinhang and Zhao, Xiangyu and Xu, Jiaxing and Zhang, Yong and Xing, Chunxiao. 2023. IMF: Interactive multimodal fusion model for link prediction. *Proceedings of the ACM Web Conference 2023*,pages:2572-2580.

Liang, Ke and Meng, Lingyuan and Liu, Meng and Liu, Yue and Tu, Wenxuan and Wang, Siwei and Zhou, Sihang and Liu, Xinwang and Sun, Fuchun and He, Kunlun. 2024. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,pages:9456-9478,IEEE.

Liu, Ye and Li, Hui and Garcia-Duran, Alberto and Niepert, Mathias and Onoro-Rubio, Daniel and Rosenblum, David S. 2019. MMKG: multi-modal knowledge graphs. *The semantic web: 16th international conference, ESWC 2019, portorož, Slovenia, June 2–6, 2019, proceedings 16*,pages:459-474,Springer.

Lu, Xinyu and Wang, Lifang and Jiang, Zejun and He, Shichang and Liu, Shizhong. 2022. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence*,pages:1-18,Springer.

Mousselly-Sergieh, Hatem and Botschen, Teresa and Gurevych, Iryna and Roth, Stefan. 2018. A multimodal translation-based approach for knowledge graph representation learning. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*,pages:225-234.

Niu, Guanglin and Li, Bo and Zhang, Yongfei and Pu, Shiliang. 2022. CAKE: A scalable commonsense-aware framework for multi-view knowledge graph completion. *arXiv preprint arXiv:2202.13785.*

Paszke, A. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703.*

Peng, Zhiliang and Dong, Li and Bao, Hangbo and Ye, Qixiang and Wei, Furu. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366.*

Shu, Yulou and Li, Wengen and Wang, Jiaqi and Zhang, Yichao and Guan, Jihong and Zhou, Shuigeng. 2025. C2RS: Multimodal Knowledge Graph Completion with Cross-Modal Consistency and Relation Semantics. *IEEE Transactions on Artificial Intelligence*,IEEE.

Sun, Rui and Cao, Xuezhi and Zhao, Yan and Wan, Junchen and Zhou, Kun and Zhang, Fuzheng and Wang, Zhongyuan and Zheng, Kai. 2020. Multi-modal knowledge graphs for recommender systems. *Proceedings of the 29th ACM international conference on information & knowledge management*,pages:1405-1414.

Sun, Zhiqing and Deng, Zhi-Hong and Nie, Jian-Yun and Tang, Jian. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197.*

Tang, Yun and Huang, Jing and Wang, Guangtao and He, Xiaodong and Zhou, Bowen. 2019. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. *arXiv preprint arXiv:1911.04910.*

Trouillon, Théo and Welbl, Johannes and Riedel, Sebastian and Gaussier, Éric and Bouchard, Guillaume. 2016. Complex embeddings for simple link prediction. *International conference on machine learning*,pages:2071-2080.PMLR.

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia. 2017. Attention is all you need. *Advances in neural information processing systems*,30.

Vrandečić, Denny and Krötzsch, Markus. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*,57(10):78-85.ACM New York, NY, USA.

Wang, Liang and Zhao, Wei and Wei, Zhuoyu and Liu, Jingming. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. *arXiv preprint arXiv:2203.02167.*

Wang, Meng and Wang, Sen and Yang, Han and Zhang, Zheng and Chen, Xi and Qi, Guilin. 2021. Is visual context really helpful for knowledge graph? A representation learning perspective. *Proceedings of the 29th ACM International Conference on Multimedia*,pages:2735-2743.

Wang, Xin and Meng, Benyuan and Chen, Hong and Meng, Yuan and Lv, Ke and Zhu, Wenwu. 2023. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. *Proceedings of the 31st ACM international conference on multimedia*,pages:2391-2399.

Wang, Zikang and Li, Linjing and Li, Qiudan and Zeng, Daniel. 2019. Multimodal data enhanced representation learning for knowledge graphs. *2019 International Joint Conference on Neural Networks (IJCNN)*,pages:1-8.IEEE.

Wilbur, W John and Sirotkin, Karl. 1992. The automatic identification of stop words. *Journal of information science*,18(1): 45-55.Sage Publications Sage CA: Thousand Oaks, CA.

Xie, Ruobing and Liu, Zhiyuan and Luan, Huanbo and Sun, Maosong. 2016. Image-embodied knowledge representation learning. *arXiv preprint arXiv:1609.07028.*

Xu, Derong and Xu, Tong and Wu, Shiwei and Zhou, Jingbo and Chen, Enhong. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. *Proceedings of the 30th ACM international conference on multimedia*,pages:3857-3866.

Yang, Bishan and Yih, Wen-tau and He, Xiaodong and Gao, Jianfeng and Deng, Li. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575.*

Zhang, Yichi and Chen, Mingyang and Zhang, Wen. 2023. Modality-aware negative sampling for multi-modal knowledge graph embedding. *2023 International Joint Conference on Neural Networks (IJCNN)*, pages:1-8.IEEE.

Zhang, Yichi and Chen, Zhuo and Guo, Lingbing and Xu, Yajing and Hu, Binbin and Liu, Ziqi and Chen, Huajun and Zhang, Wen. 2024. Mygo: Discrete modality information as fine-grained tokens for multi-modal knowledge graph completion. *arXiv preprint arXiv:2404.09468.*

Zhang, Yichi and Chen, Zhuo and Guo, Lingbing and Xu, Yajing and Hu, Binbin and Liu, Ziqi and Zhang, Wen and Chen, Huajun. 2024. Native: Multi-modal knowledge graph completion in the wild. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*,pages:91-101.

Zhang, Yichi and Chen, Zhuo and Liang, Lei and Chen, Huajun and Zhang, Wen. 2024. Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. *arXiv preprint arXiv:2402.15444.*

Zhang, Yichi and Zhang, Wen. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *arXiv preprint arXiv:2209.07084.*

Zhao, Yu and Cai, Xiangrui and Wu, Yike and Zhang, Haiwei and Zhang, Ying and Zhao, Guoqing and Jiang, Ning. 2022. Mose: Modality split and ensemble for multimodal knowledge graph completion. *arXiv preprint arXiv:2210.08821.*