Differentiated Vision: Unveiling Entity-Specific Visual Modality Requirements for Multimodal Knowledge Graph

Minghang Liu[1,2]* **Yinghan Shen**[1]* **Zihe Huang**[1,2]*
Yuanzhuo Wang[1,✉] **Xuhui Jiang**[3,4] **Huawei Shen**[1]

[1] State Key Laboratory of AI Safety, Institute of Computing Technology, CAS
[2] University of Chinese Academy of Sciences [3] DataArc Tech Ltd.
[4] IDEA Research, International Digital Economy Academy
liuminghang23s@ict.ac.cn, ✉ wangyuanzhuo@ict.ac.cn

Abstract

Multimodal Knowledge Graphs (MMKGs) enhance knowledge representations by integrating structural and multimodal information of entities. Recently, MMKGs have proven effective in tasks such as information retrieval, knowledge discovery, and question answering. Current methods typically utilize pre-trained visual encoders to extract features from images associated with each entity, emphasizing complex cross-modal interactions. However, these approaches often overlook the varying relevance of visual information across entities. Specifically, not all entities benefit from visual data, and not all associated images are pertinent, with irrelevant images introducing noise and potentially degrading model performance. To address these issues, we propose the **D**ifferentiated **V**ision for **M**ultimodal **K**nowledge **G**raphs (**DVMKG**) model. DVMKG evaluates the necessity of visual modality for each entity based on its intrinsic attributes and assesses image quality through representativeness and diversity. Leveraging these metrics, DVMKG dynamically adjusts the influence of visual data during feature integration, tailoring it to the specific needs of different entity types. Extensive experiments on multiple benchmark datasets confirm the effectiveness of DVMKG, demonstrating significant improvements over existing methods.

1 Introduction

Multimodal Knowledge Graphs (MMKGs) extend traditional knowledge graphs by integrating multimodal data, such as images and text, to overcome the limitations of incomplete or ambiguous entity representations that arise from relying solely on textual information(Liu et al., 2019). By incorporating visual data , MMKGs generate richer entity embeddings, improving performance in downstream tasks, including knowledge graph completion (Xu et al., 2024b; Cao et al., 2022; Liang et al., 2023b), entity alignment(Li et al., 2023a; Lin et al., 2022) and information retrieval(Dietz et al., 2018; Yang, 2020).

Existing MMKG approaches focus on the encoding and fusion of different modalities (Chen et al., 2024b): late fusion (Lu et al., 2022; Wang et al., 2022; Li et al., 2022b) combine modality features just before output, and early fusion (Fang et al., 2022; Wei et al., 2023) integrate features earlier to foster deeper cross-modal interactions. Recent studies (Shang et al., 2024; Zhang et al., 2024c) address the challenges of integrating visual information by reducing noise through link-specific image relevance and improving image-text matching for long-tail entities. However, these methods assume that the visual modality uniformly enhances performance, neglecting to evaluate its necessity for entities, thus lacking a systematic approach to assess its entity-specific benefits.

To effectively harness visual knowledge in MMKGs, two critical aspects must be considered: entity-level necessity and image-level selection. As shown in Figure 1(a), experiments reveal that randomly removing visual data—either by excluding all images for 10% of entities or 10% of images across all entities—can unexpectedly improve model performance. This indicates that not all entities benefit from visual features. For example, as illustrated in Figure 1(b), entities with distinct visual characteristics, such as *FC_Schalke_04* gain from visual data, which provides additional context for predictive accuracy. Conversely, abstract entities like *Norwegian_ language* often lack relevant visual cues, and including such data may fail to capture their conceptual depth, potentially degrading performance. At the image level, when multiple relevant images are available, selecting those that meaningfully enhance the entity representation is essential. As illustrated with *FC_Schalke_04*, im-

*Equal contribution.
✉Corresponding author: ✉ wangyuanzhuo@ict.ac.cn

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20170–20183
November 4-9, 2025 ©2025 Association for Computational Linguistics

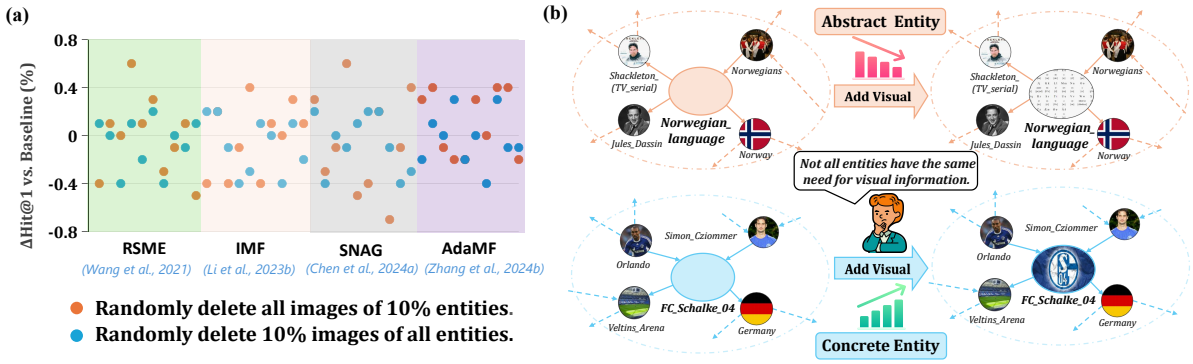

Figure 1: (a) Experimental results of randomly removing part of the visual modality. The y-axis shows the relative change compared to the model with all images (baseline, y=0). Positive values indicate improvement, while negative values indicate degradation. (b) Visual data improves performance for concrete entities (e.g., "*FC_Schalke_04*") but diminishes it for abstract ones (e.g., "*Norwegian_language*").

ages are evaluated based on representativeness and diversity. The team logo, with its high representativeness score, captures the club's identity, while diverse images like team photos and stadium views provide broader perspectives, supporting a well-rounded understanding. Balancing representativeness and diversity prevents overfitting to specific visual cues and enhances the richness of entity representations in the knowledge graph.

To address the above issues, we introduce the **D**ifferentiated **V**ision for **M**ultimodal **K**nowledge **G**raphs (DVMKG) Model, which classifies entities as abstract or concrete based on inherent properties and assigns a visualizability score to determine the necessity of visual data for each entity. At the entity level, a large language model (LLM) assigns a visualizability score to each entity, determining its need for visual data. For image selection, we leverage object detection and Pretrained Vision-Language Model (PVLM) to evaluate semantic alignment, ensuring visual accuracy while promoting diversity to reflect various entity facets. A multi-head attention mechanism dynamically adjusts visual embeddings according to entity-specific requirements, and inter-modal contrastive learning aligns representations across modalities. Finally, a weighted fusion layer integrates these embeddings to optimize the multimodal representation. Our main contributions are summarized as follows:

- We introduce the DVMKG model, which categorizes entities as abstract or concrete and assigns visualizability scores to prioritize those benefiting most from visual data, while optimizing image selection using representativeness and diversity metrics.
- We employ a multi-head attention mechanism and inter-modal contrastive learning to dynamically adjust and align visual embeddings, fused via a weighted mutan layer for a comprehensive multimodal representation.
- We validate DVMKG on three widely used datasets, demonstrating its effectiveness in enhancing MMKG representations through entity-specific visual integration.

2 Assessing Visual Modality in MMKG

In MMKG, visual modality can enhance knowledge graph representation, yet its necessity varies across entities. Even for entities that benefit from visual information, the most suitable images differ. This section explores how to identify entities that require visual information and select images to represent them, addressing this challenge at two levels: entity level and image level. At the entity level, we assess the relevance of visual data by scoring entity visualizability. At the image level, we focus on selecting images for entities to capture essential characteristics while maintaining diversity.

2.1 The definition of MMKG

Multi-modal Knowledge Graph. A MMKG can be formally defined as $G = (\mathcal{E}, \mathcal{R}, \mathcal{M}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations between entities, $\mathcal{M} = \{s, v, t\}$ is the set of multi-modal data, where s, v, t denote structural, visual, textual modality, and $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ represents the relational triples of the knowledge graph. Each triple in the graph is represented as (h, r, t, m_h, m_t), where h and t are head and tail entities, r is the relation between them, and m_h,

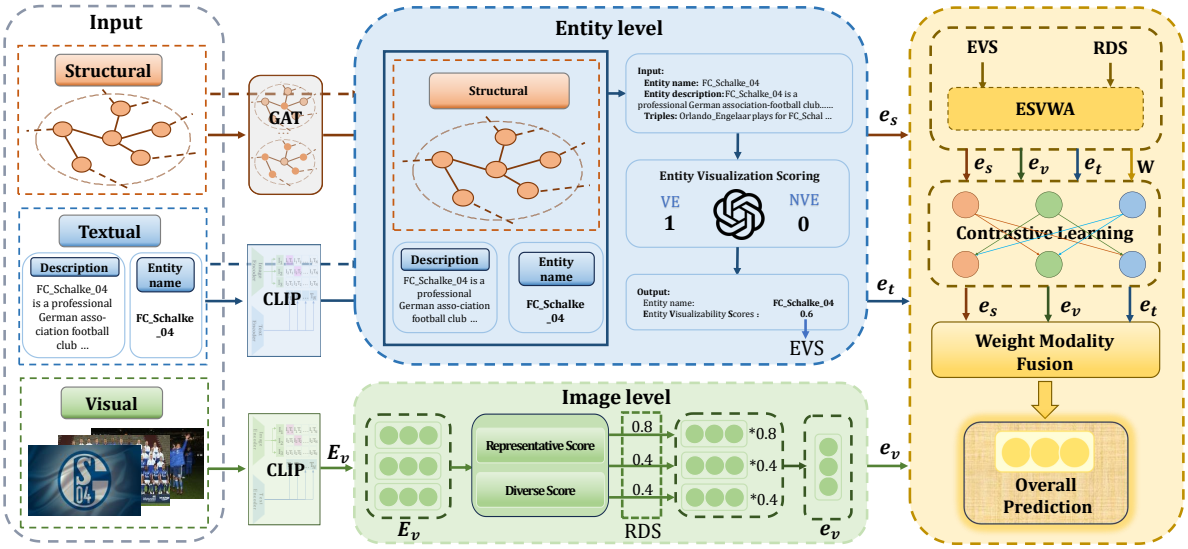

Figure 2: Overview of the DVMKG. At the Entity Level, LLM generate an entity visualizability score (EVS). At the Image Level, vision-language models assess representativeness and diversity, which are combined into a representativeness-diversity score (RDS) to adjust visual embeddings. Visual (e^v), structural (e^s), and textual (e^t) embeddings are fused through contrastive learning and Weight Modality Fusion to provide the final prediction.

m_t are the multi-modal information attached to the respective entities.

2.2 Assessing Entity Visualizability

Not all entities in the MMKG benefit from the visual information. Concrete entities, such as person type entities, are readily associated with distinct visual forms, whereas abstract entities, like language or culture type entities, are difficult to be described by visual attributes. We introduce the following definition based on their degree of abstraction or concreteness:

Non-visualizable Entity (NVE): Entities lacking a direct visual representation due to their abstract nature.

Visualizable Entity (VE): Entities with a clear, concrete visual representation.

Recognizing that concreteness and abstraction exist on a continuum rather than as binary opposites(Villani et al., 2022). To quantify this, we introduce the **Entity Visualizability Score (EVS)**, a continuous measure ranging from 0 to 1, where 0 denotes highly abstract, non-visualizable entities, and 1 denotes highly visualizable entities.

To assign EVS to entities, we leverage a LLM guided by prompt engineering strategies, detailed in Appendix B. The LLM's extensive prior knowledge improves the accuracy of visualizability estimation. Given that LLMs struggle with abstract concepts compared to concrete ones (Liao et al., 2023), we enhance their capability by integrating two information sources: (1) textual entity descriptions and (2) structural information from entity triples. The structured triples are transformed into natural language sentences (e.g., the triple"<*Orlando_Engelaar, current_team, FC_Schalke_04*>" is converted to the sentence "Orlando Engelaar plays for FC Schalke 04"), thereby enriching the contextual cues available to the model.

2.3 Evaluating the Representativeness and Diversity of Visual Information

In the previous section, we identified entities that benefit from visual modality at the entity level. However, for entities associated with multiple images, selecting the most effective images for representation is essential. To address this, we evaluate visual modality at the image level using two metrics: **Representative Image Score** and **Diversity Image Score** (RDS).

2.3.1 Representative Image Score

Concrete entities are often associated with a diverse set of images. For instance, a public figure may appear in various contexts, such as formal events or casual moments, each reflecting different facets of their identity. To quantify the representativeness of an image I_i for a given entity e, we propose a scoring mechanism that integrates two components:

a PVLM and the object detection component. The method ensures that the selected image aligns with the entity's name and resembles its typical visual representation.

Let $I = \{I_1, I_2, \ldots, I_n\}$ denote the set of images related to entity e. The representative image score $S_{Rep}(I_i, e)$ for each image I_i is computed as a weighted sum:

$$S_{Rep}(I_i, e) = \alpha \cdot S_{PVLM_R} + (1 - \alpha) \cdot S_{OD}, \quad (1)$$

where $\alpha \in [0, 1]$ is a weighting parameter balancing the contributions of the two components.

PVLM Component: The PVLM evaluates the alignment between an entity's name $name(e)$ and the image I_i. It generates a similarity score, transformed via the Sigmoid function σ:

$$S_{\text{PVLM}_R}(I_i, e) = \sigma(\text{PVLM}(I_i, name(e))). \quad (2)$$

Here, $\text{PVLM}_R(I_i, name(e))$ denotes the raw similarity score output by the PVLM.

Object Detection Component: A typical image I_{typ} is selected to encapsulate the entity's key visual characteristics (details are provided in Appendix C.1). The object detection component measures the visual similarity between $I_i$and the typical image I_{typ} of an entity e:

$$S_{OD}(I_i, I_{typ}) = \max_{o_i^j \in \text{Obj}(I_i)} \left(\cos\left(o_j, I_{typ}\right)\right), \quad (3)$$

where $\text{Obj}(I_i)$ is the set of detected objects in image I_i, o_j represents the j-th detected object. The maximum similarity ensures that the most representative object in I_i is considered.

2.3.2 Diverse Image Score.

While representativeness focuses on core characteristics, diversity ensures that the chosen set of images reflects a broader range of the entity's attributes. The diversity image score S_{Div} for an image I_i of entity e evaluates the image's alignment with the entity's textual description and its uniqueness compared to other images:

$$S_{Div}(I_i, e) = \beta \cdot S_{\text{PVLM}_D}(I_i, e) + (1-\beta) \cdot S_{Un}(I_i, I), \quad (4)$$

where $\beta \in [0, 1]$ balances the contributions of the PVLM and uniqueness components.

PVLM Component for Diversity: Unlike the representativeness score, this component aligns the image with keywords extracted from the entity's textual description t. For example, entity *FC_Schalke_04* described as "a professional German football and multi-sports club from Gelsenkirchen," keywords such as "football" and "German" guide the selection of diverse images, such as football-related scenes and German culture. Let $key(t) = (k_1, k_2, \ldots, k_m)$ denote the set of keywords from the textual description t. The score is:

$$S_{\text{PVLM}_D} = \sum_{k_j \in \text{key}(t)} \sigma(\text{PVLM}(I_i, k_j)). \quad (5)$$

This aggregates the similarity scores between I_i and each keyword, capturing a wider range of visual attributes.

Uniqueness Component: The uniqueness score $S_{Un}(I_i, I)$ ensures that I_i provides distinct information by comparing it to other images in I:

$$S_{Un} = \frac{1}{n-1} \sum_{j \neq i} (1 - \cos(V(I_i), V(I_j))), \quad (6)$$

where $V(I_i)$ represents the feature vector of I_i extracted by the PVLM, n is the number of images of e.

By combining S_{Rep} and S_{Div}, we select images that both align with the entity's core identity and capture its diverse attributes. This dual-scoring framework enhances the visual modality's role in MMKGs, providing a richer and more comprehensive entity representation.

3 Methodology

To effectively leverage the visual modality needs of different entities, we propose the DVMKG framework. The EVS and RDS serve as key inputs to our framework, integrating entity-specific visual demands into the Multimodal Knowledge Graph Completion (MKGC) task. As shown in Figure 2, we first introduce the pre-trained encoders for different modalities. We then detail the Entity-Specific Visual Weight Adjustment (ES-VWA) module, inter-modal contrastive learning, and the Weighted Modality Fusion mechanism.

3.1 Multiple Modality Feature Encoding

We use pretrained encoders to encode each modality in the MMKG, complementary representations for subsequent fusion and learning. The embeddings obtained from the structural, visual, and textual modalities are denoted as $\mathbf{E^s}$, $\mathbf{E^v}$, and $\mathbf{E^t}$, respectively. The detailed process is described in Appendix C.5.

3.2 Entity-Specific Visual Weight Adjustment

This ESVWA module dynamically adjusts the contribution of visual embeddings based on each entity's specific visual modality requirements and the quality of the available visual data. This adaptive mechanism enhances the entity's representation during multimodal fusion.

The EVS and the RDS are first stacked along the feature dimension to consolidate them into a unified tensor representation $\mathbf{I} \in \mathbb{R}^{2\times d}$, with d representing the embedding dimension. Next, the stacked tensor is projected into a lower-dimensional space using a linear transformation:

$$\mathbf{I}_{vis} = W_1 \cdot \mathbf{I} + \mathbf{b}_1, \tag{7}$$

where $W_1 \in \mathbb{R}^{d'\times 2}$ and $\mathbf{b}_1 \in \mathbb{R}^{d'}$ are learnable parameters.

To further refine the visual embeddings, a self-attention mechanism is applied. The input to the self-attention layer is the projected tensor $\mathbf{I}_{vis}$ and the output of the self-attention mechanism is:

$$\begin{aligned} \mathbf{A} &= \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d'}}\right)\mathbf{V}, \\ \mathbf{Q} = \mathbf{W}_Q \cdot \mathbf{I}_{\text{vis}}, \mathbf{K} &= \mathbf{W}_K \cdot \mathbf{I}_{\text{vis}}, \mathbf{V} = \mathbf{W}_V \cdot \mathbf{I}_{\text{vis}}. \end{aligned} \tag{8}$$

Here, $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are the query, key, and value matrices derived from $\mathbf{I}_{vis}$. $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d'}$ are learnable parameters of the attention layer. Finally, $\mathbf{A}$ are applied to the initial visual modality embeddings $\mathbf{E}^v$ refining the visual representation:

$$\mathbf{E}'_{\mathbf{v}} = \mathbf{A} \odot \mathbf{E}_{\mathbf{v}}, \tag{9}$$

where $\odot$ denotes element-wise multiplication, ensuring that the contribution of each dimension in the visual embedding is scaled according to its dynamically computed importance.

3.3 Inter-modal Contrastive Learning

To ensure consistency across modalities and reduce interference, we employ inter-modal contrastive learning to align entity representations. The contrastive loss is defined as:

$$\mathcal{L}_{CL} = -\sum_{p,q\in\mathcal{M}}\sum_{i\in I}\log\frac{\exp\left(\cos(\boldsymbol{e}_i^p, \boldsymbol{e}_i^q)/\tau\right)}{\sum_{j\in A(i)}\exp\left(\cos(\boldsymbol{e}_i^p, \boldsymbol{e}_i^q)/\tau\right)}, \tag{10}$$

where $\mathcal{M} = \{s, v, t\}$ denotes the set of multimodal data. $\boldsymbol{e}_i^p$ and $\boldsymbol{e}_i^q$ are embeddings of entity i from modalities p and q, respectively, and τ is a temperature parameter that controls similarity sensitivity.

3.4 Weight Modality Fusion

Inspired by the Mutan model (Ben-Younes et al., 2017), we introduce a weighted modality fusion layer that integrates modality-specific embeddings with entity-specific visual importance:

$$\text{WMF}(\mathbf{E}^{\mathbf{s}}, \mathbf{E}^{\mathbf{v}}, \mathbf{E}^{\mathbf{t}}) = (W_s\mathbf{E}^{\mathbf{s}}) \odot (W_v\mathbf{E}^{\mathbf{v}})) \odot \left(W_t\mathbf{E}^{\mathbf{t}}\right), \tag{11}$$

where W_s, W_v, and W_t are the learnable weight matrices, and $\odot$ denotes element-wise multiplication. The multimodal embedding $\mathbf{E}^{\mathbf{mm}}$ is:

$$\mathbf{E}^{mm} = \text{ReLU}\left(\text{WMF}\left(\mathbf{E}^{\mathbf{s}}, \mathbf{E}'_{\mathbf{v}}, \mathbf{E}^{\mathbf{t}}\right)\right). \tag{12}$$

3.5 Training

To assess the difference between predicted values $\hat{y}$ and true labels y, we employ the cross-entropy loss function, which measures the discrepancy between actual values and the predicted outputs. Specifically, the modality-specific cross-entropy loss $\mathcal{L}_M$ is computed as follows:

$$\begin{aligned} \mathcal{L}_M = -\frac{w^p}{|\mathcal{E}|}\sum_{p\in\mathcal{M}}\sum_{i=1}^{|\mathcal{E}|}(y_{i,p}\cdot\log(\hat{y}_{i,p})) \\ + (1-y_{i,p})\cdot\log(1-\hat{y}_{i,p}), \end{aligned} \tag{13}$$

where w^p is the weight for modality p, $\mathcal{M} = \{s, t, v, mm\}$ denotes the set of modalities, and $\mathcal{E}$ represents the set of entities. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_M + \lambda\mathcal{L}_{CL}, \tag{14}$$

where λ is a hyperparameter that balances the contribution of the two loss terms.

4 Experiments

4.1 Experimental Basic Setup

We evaluate our model on three widely adopted multimodal knowledge graph (MMKG) datasets: DB15K(Liu et al., 2019), YAGO15K(Liu et al., 2019) and FB15K-237(Mousselly-Sergieh et al., 2018). To assess the effectiveness of our proposed DVMKG model, we compare it with representative baselines from both unimodal and multimodal knowledge graph completion methods. All implementation details, including experimental settings and hyperparameter choices, are provided in Appendix C.

4.2 Main Experiment

As shown in Table 1, DVMKG consistently achieves state-of-the-art results in the MKGC task,

	DB15K				YAGO15K				FB15K-237			
	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR
Unimodal methods												
TransE (Bordes et al., 2013)	0.128	0.315	0.471	0.249	0.185	0.271	0.381	0.201	0.198	0.376	0.441	0.294
DistMult (Yang et al., 2015)	0.148	0.263	0.396	0.230	0.215	0.316	0.438	0.291	0.199	0.301	0.446	0.241
ConvE (Dettmers et al., 2018)	0.219	0.328	0.507	0.294	0.168	0.261	0.426	0.267	0.237	0.356	0.501	0.325
TuckER(Balažević et al., 2019)	0.233	0.380	0.480	0.330	0.183	0.276	0.457	0.281	0.261	0.394	0.536	0.353
KC-GenRe (Wang et al., 2024)	0.240	0.385	0.483	0.336	0.188	0.280	0.460	0.282	0.256	0.387	0.533	0.329
MPIKGC (Xu et al., 2024a)	0.227	0.373	0.475	0.321	0.180	0.275	0.459	0.280	0.267	0.395	0.543	0.358
Multimodal methods												
IKRL (Xie et al., 2016)	0.141	0.349	0.491	0.268	0.162	0.247	0.346	0.251	0.194	0.284	0.458	0.309
TransAE (Wang et al., 2019)	0.213	0.312	0.412	0.281	0.177	0.262	0.359	0.268	0.199	0.317	0.463	0.315
RSME (Wang et al., 2021)	0.241	0.321	0.402	0.298	0.233	0.319	0.414	0.305	0.242	0.344	0.467	0.331
VISTA (Lee et al., 2023)	0.247	0.350	0.469	0.308	0.280	0.411	0.485	0.347	0.240	0.343	0.468	0.335
IMF (Li et al., 2023b)	0.251	0.362	0.482	0.329	0.288	0.396	0.499	0.360	0.267	0.396	0.551	0.360
SNAG (Chen et al., 2024a)	0.247	0.407	0.528	0.345	0.310	0.423	0.519	0.384	0.234	0.365	0.520	0.329
AdaMF (Zhang et al., 2024b)	0.253	0.411	0.529	0.351	0.307	0.426	0.525	0.385	0.268	0.397	0.539	0.359
DVMKG(Ours)	**0.307***	**0.424***	**0.538***	**0.386***	**0.360***	**0.470***	**0.534***	**0.419***	**0.285***	**0.413***	**0.568***	**0.379***

Table 1: Experimental results on datasets from MMKG. The most competitive baseline results are underlined and the best outcomes are highlighted in **bold**. The superscript * denotes statistically significant improvement ($p < 0.05$).

outperforming all unimodal and multimodal baselines across the three benchmark datasets. This underscores the efficacy of our approach in leveraging entity-specific visual information. Among the unimodal methods, MPIKGC and KC-GenRe , which utilize large language models (LLMs), exhibit strong performance relative to other unimodal techniques. Despite this advancement within unimodal strategies, these LLM-enhanced unimodal models are generally surpassed by several multimodal approaches. This suggests that while advanced textual understanding is beneficial, integrating visual modalities provides complementary information that significantly enhances overall MKGC performance.

4.3 Ablation Experiment

4.3.1 Impact of Scoring Modules

To assess the contribution of the two scoring modules that guide the adjustment of the visual representations of the entities, we performed ablation experiments under three conditions: (1) *w/o EVS*: removing the Entity Visualizability Score; (2) *w/o RDS*: removing the Representative Image Score and Diversity Image Score; (3) *w/o EVS+RDS*: removing both EVS and RDS simultaneously.

As shown in Table 2, the ablation studies confirm the effectiveness of EVS, RDS, and their combination in refining the visual modality. The elimination of either scoring module degrades overall performance. Notably, the entity visualizability score exerted a greater influence compared to the representativeness and diversity scores, indicating that for entities with lower-quality visual data, adjusting the visual modality is likely more critical than determining the relative weight of the image in the final embedding.

Model	Hits@1	Hits@3	Hits@10	MRR
DVMKG	**0.360**	**0.447**	**0.534**	**0.419**
-w/o EVS	0.345	0.430	0.521	0.406
-w/o RDS	0.347	0.432	0.525	0.408
-w/o EVS+RDS	0.341	0.427	0.517	0.401

Table 2: Ablation study on YAGO15K.

Model	MSE	MAE
GPT-4o (Achiam et al., 2023)	0.0855	0.2126
DeepSeek-v3 (Liu et al., 2024)	0.0930	0.2354
GLM-4-plus (GLM et al., 2024)	0.0820	0.2124
Gemini-2.0-pro (Team et al., 2025)	0.0810	0.2126
Qwen2.5-Max (Yang et al., 2024)	**0.0620**	**0.1943**
-w/o triples	0.0875	0.2270
-w/o text	0.1211	0.2760
-w/o text+triples	0.1400	0.2980

Table 3: Scoring entities based on Visualizability.

4.3.2 Ablation Study on LLM-Based Visualizability Scoring

Evaluating LLM Performance for Visualizability Scoring. To evaluate the performance of different LLMs and validate the effectiveness of each component in our LLM-based scoring mechanism, we conducted an ablation study. We manually annotated 10% of the entities with visualizability scores for the test set. The comparison of different LLMs in Table 3 demonstrates varying levels of performance in visualizability scoring, with Qwen2.5-

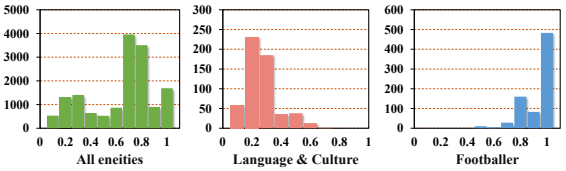

Figure 3: EVS and its domain-specific distributions in Language&Culture and Footballer on YAGO15K.

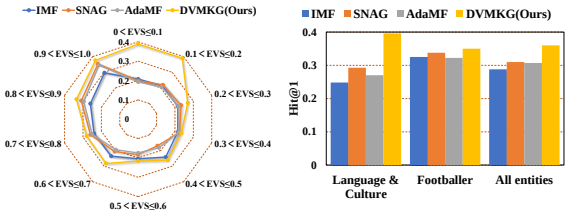

Figure 4: Hit@1 of DVMKG and baselines across EVS intervals and Domain-Specific on YAGO15K.

Max achieving the best results. We also compares different LLM configurations in this task. Only using entities provides baseline performance, while adding structural and textual descriptions progressively improves accuracy. These results highlight that integrating structural and textual information significantly enhances the LLM's accuracy in assessing visualizability, providing a reliable measure of entity abstraction and concreteness.

Analysis of EVS Distribution and Its Impact. The overall distribution and the domain-specific distributions of EVS are illustrated in Figure 3. Overall distribution reveals significant variations in EVS across different entities, indicating that the LLM effectively distinguishes the visualizability needs of entities. In the domain of "Language & Culture," the EVS values are generally lower due to their abstract nature, reflecting that entities in this domain rely less on visual modality. In contrast, entities in the "Footballer" domain exhibit higher EVS values.

As shown in Figure 4, the radar chart illustrates Hit@1 various EVS ranges. Notably, DVMKG consistently outperforms the other models, particularly in the lower EVS ranges (0-0.2). This highlights DVMKG's ability to handle entities with low visualizability effectively by dynamically adjusting the weight of the visual modality. As EVS increases, the performance gap narrows, indicating that most models benefit from highly visualizable entities. Nevertheless, DVMKG maintains a advantage across all EVS ranges.

The bar chart further validates the superiority of the DVMKG model in specific domains. In

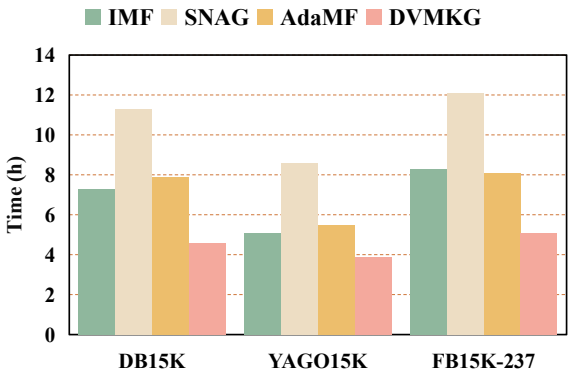

Figure 5: The results of the efficiency experiment.

the Language & Culture domain, DVMKG significantly outperforms other models, demonstrating its ability to effectively handle the visual information needs of abstract entities. In the Footballer domain, DVMKG also excels, further proving its robustness in processing concrete entities.

Effectiveness of components in representativeness and diversity scoring is presented in Appendix D and E.

4.4 Efficency Analysis

In the efficiency experiment, we compared the training times of IMF, SNAG, AdaMF, and our proposed DVMKG model. As shown in Figure 5, DVMKG achieved the shortest training time in all datasets. The IMF model employs a two-stage fusion, maintaining moderate efficiency. The Transformer-based architecture used in SNAG results in the longest training time due to its complexity. AdaMF employs adaptive modality weights but remains less efficient than our model. DVMKG integrates LLM and PVLM to enhance visual understanding capabilities. While leveraging these advanced models typically increases computational complexity, DVMKG achieves a balance between performance and efficiency. Although an LLM is part of the architecture, its primary role is confined to a straightforward scoring task. Furthermore, DVMKG's design incorporates a simpler model architecture and exhibits faster convergence speed. This rapid convergence is facilitated by effective initialization. By doing so, DVMKG minimizes unnecessary computational load, focusing resources on entities where visual information is most relevant.

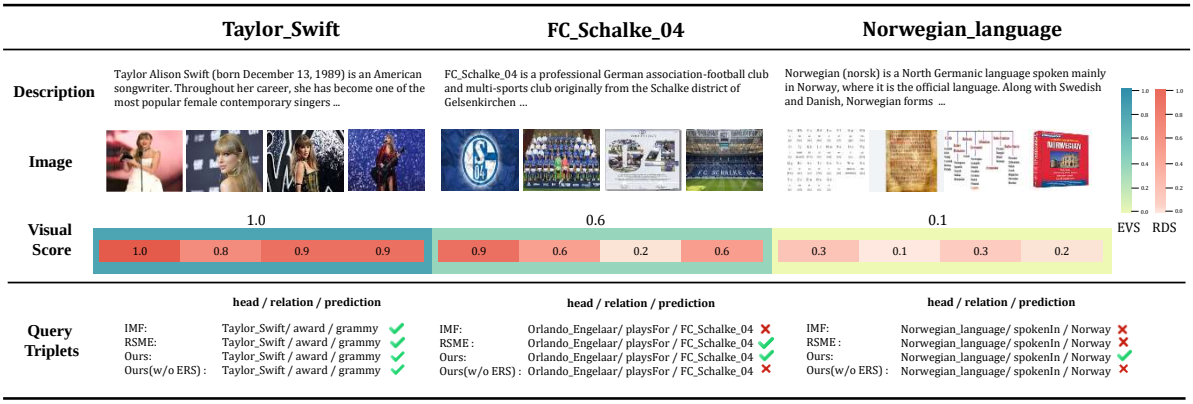

Figure 6: Case study of visual modality evaluation for different entities. This figure presents a comparative analysis of the visual modality evaluation for three different entities from the dataset: "*Taylor_Swift*," "*FC_Schalke_04*," and "*Norwegian_language*." Each entity is characterized by its description, images, visual score, and corresponding knowledge graph query triplets.

4.5 Case study

To intuitively analyze our experimental results, we conducted case studies by selecting one entity each from the high, intermediate and low entity visualizability score categories. We compared two models: IMF, which employs a parameterization to adjust the contribution of each modality, and RSME, which selectively integrates relevant visual data to optimize entity representations by filtering out less impactful information. We also included an ablation study, referred to as "ours(w/o ERS)," where both EVS and RDS were removed to eliminate their combined effect on the visual modality. As shwon in Figure 6, we can further notice through the case:

- "*Taylor_Swift*"(High EVS): As a highly visualizable entity, Taylor Swift's visual information effectively complements other modalities, requiring minimal adjustments. In this case, all models successfully made the correct prediction.

- "*FC_Schalke_04*"(Intermediate EVS): In this case, its associated images include those that are representative (e.g., team logos and group photos) and others that are less relevant. Our DVMKG and RMSE model reduced the weight of the visual modality and irrelevant images to minimize interference, which led to successful predictions. However, the IMF and "DVMKG-w/o ERS" failed due to limited modality adjustment.

- "*Norwegia_language*"(Low EVS): DVMKG dynamically lowers the visual modality's weight, relying on textual and structural data for correct predictions. RSME fails by filtering low-scoring images without accounting for entity-specific visual needs.

These case studies demonstrate the effectiveness of our model in dynamically adjusting the visual modality based on the visual score of each entity. By selectively reducing or emphasizing visual information, our approach optimally integrates visual data to enhance the representation of MMKGs.

5 Conclusion

In this paper, we propose the DVMKG model to address the critical challenge of effectively integrating visual information in MMKGs, that not all entities benefit equally from visual data and not all images are pertinent. At the entity level, DVMKG evaluates the degree of visualizability for each entity. At the image level, DVMKG assesses the quality of associated images through comprehensive representativeness and diversity scores. Based on these scores, DVMKG dynamically adjusts the influence of visual information during the feature integration process, employing techniques such as multi-head attention for refined visual embeddings and contrastive learning for cross-modal alignment. Our extensive experiments conducted on multiple benchmark datasets robustly demonstrate DVMKG's superiority over existing methods. The results confirm that an entity-specific approach to visual modality integration can significantly enhance the representational capabilities of Multimodal Knowledge Graphs.

6 Limitations

Although the DVMKG model demonstrates significant effectiveness in its current application, an exciting direction for future development would be to extend the DVMKG framework to incorporate other data types, like audio or interactive media, which could further enhance the depth and versatility of multimodal knowledge graph representations. Additionally, a valuable area for future investigation is to explore the application of DVMKG's entity-specific visual assessment principles more widely. This includes potential uses in a broader range of image-enhanced natural language processing, such as multimodal question answering, topic detection, and document classification.

7 Ethics Statement

To the best of our knowledge, this work does not involve any discrimination, social bias, or private data. All the datasets are constructed from open source KGs such as Wikidata, YAGO, and DBpedia.

Acknowledgments

Thanks to all reviewers, their reviews are important for this research. This work is supported by the National Natural Science Foundation of China (No.62172393), Major Public Welfare Project of Henan Province (No.201300311200), and Henan provincial key research and development program (No. 241111211900)

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099.

Ivana Balažević, Carl Allen, and Timothy Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

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

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems*, 35:39090–39102.

Zhuo Chen, Yin Fang, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Huajun Chen, and Wen Zhang. 2024a. The power of noise: Toward a unified multi-modal knowledge graph representation framework. *arXiv preprint arXiv:2403.06832*.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, and 1 others. 2024b. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

P Kingma Diederik. 2014. Adam: A method for stochastic optimization. *(No Title)*.

Laura Dietz, Alexander Kotov, and Edgar Meij. 2018. Utilizing knowledge graphs for text-centric information retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1387–1390.

Quan Fang, Xiaowei Zhang, Jun Hu, Xian Wu, and Changsheng Xu. 2022. Contrastive multi-modal knowledge graph representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8983–8996.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Qian Li, Cheng Ji, Shu Guo, Zhaoji Liang, Lihong Wang, and Jianxin Li. 2023a. Multi-modal knowledge graph transformer framework for multi-modal entity alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 987–999.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023b. Imf: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, pages 2572–2580.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Yancong Li, Xiaoming Zhang, Fang Wang, Bo Zhang, and Feiran Huang. 2022b. Fusing visual and textual content for knowledge graph embedding via dual-track model. *Applied Soft Computing*, 128:109524.

Ke Liang, Sihang Zhou, Yue Liu, Lingyuan Meng, Meng Liu, and Xinwang Liu. 2023a. Structure guided multi-modal pre-trained transformer for knowledge graph reasoning. *arXiv preprint arXiv:2307.03591*.

Shuang Liang, Anjie Zhu, Jiasheng Zhang, and Jie Shao. 2023b. Hyper-node relational graph attention network for multi-modal knowledge graph completion. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–21.

Jiayi Liao, Xu Chen, and Lun Du. 2023. Concept understanding in large language models: An empirical study. In *Tiny Papers @ ICLR*.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-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 16*, pages 459–474. Springer.

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

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

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6967–6977.

Bin Shang, Yinliang Zhao, Jun Liu, and Di Wang. 2024. Lafa: Multimodal knowledge graph completion with link aware fusion and aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8957–8965.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Caterina Villani, Matteo Orsoni, Luisa Lugli, Mariagrazia Benassi, and Anna M Borghi. 2022. Abstract and concrete concepts in conversation. *Scientific Reports*, 12(1):17572.

Enqiang Wang, Qing Yu, Yelin Chen, Wushouer Slamu, and Xukang Luo. 2022. Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Information Fusion*, 88:78–85.

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

Yilin Wang, Minghao Hu, Zhen Huang, Dongsheng Li, Dong Yang, and Xicheng Lu. 2024. KC-GenRe: A knowledge-constrained generative re-ranking method based on large language models for knowledge graph completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9668–9680, Torino, Italia. ELRA and ICCL.

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

Yuyang Wei, Wei Chen, Shiting Wen, An Liu, and Lei Zhao. 2023. Knowledge graph incremental embedding for unseen modalities. *Knowledge and Information Systems*, 65(9):3611–3631.

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

Derong Xu, Ziheng Zhang, Zhenxi Lin, Xian Wu, Zhihong Zhu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024a. Multi-perspective improvement of knowledge graph completion with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11956–11968, Torino, Italia. ELRA and ICCL.

Yunhui Xu, Youru Li, Muhao Xu, Zhenfeng Zhu, and Yao Zhao. 2024b. Hka: A hierarchical knowledge alignment framework for multimodal knowledge graph completion. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(8):1–19.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

Zuoxi Yang. 2020. Biomedical information retrieval incorporating knowledge graph for explainable precision medicine. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2486–2486.

Nannan Yu, Tianshuang Qiu, Feng Bi, and Aiqi Wang. 2011. Image features extraction and fusion based on joint sparse representation. *IEEE Journal of selected topics in signal processing*, 5(5):1074–1082.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. 2022. Multimodal analogical reasoning over knowledge graphs. In *The Eleventh International Conference on Learning Representations*.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 2024b. 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)*, pages 17120–17130.

Yikai Zhang, Qianyu He, Xintao Wang, Siyu Yuan, Jiaqing Liang, and Yanghua Xiao. 2024c. Light up the shadows: Enhance long-tailed entity grounding with concept-guided vision-language models. *arXiv preprint arXiv:2406.10902*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

A Related Work

A.1 MMKG Representation Learning

MMKG representation learning methods focus on integrating diverse data modalities such as text and images within knowledge graphs, enhancing the modeling and understanding of complex data structures and improving entity representations (Fang et al., 2022; Wang et al., 2022; Liang et al., 2023a). By incorporating such diverse information, these methods are able to enrich entity representations and improve reasoning in multimodal scenarios. For instance, MMKRL (Lu et al., 2022) employs attention and gating mechanisms for cross-modal interactions, resulting in more robust entity representations in noisy or incomplete data scenarios. Building on this, MMRotatH (Wei et al., 2023) facilitates the incremental integration of new modalities without requiring full model retraining, ensuring adaptability in dynamic environments. Additionally, DuMF (Li et al., 2022b) fuses visual and textual content to enrich entity representations by capturing complementary information from different modalities. MARS (Zhang et al., 2022) further enhances entity representations by incorporating multimodal data to improve analogical reasoning, thereby producing more accurate and comprehensive embeddings in knowledge graph tasks.

A.2 Visual Modality Integration in Multimodal Representation

In recent years, the integration of multimodal data has gained significant traction across various fields, as it enables models to leverage complementary

information from diverse modalities. The selection, processing, and fusion of visual modalities have become pivotal to improving representation learning. Existing research often approaches this from two perspectives: refining the image modality internally and improving cross-modal alignment between visual and textual data. For intra-Visual Modality, images are divided into common and innovation components, with joint sparse representation used to model both shared and unique features (Yu et al., 2011). TransFusion (Bai et al., 2022) further develop a soft-association mechanism using self-attention to adaptively focus on relevant visual details, handling image quality variations and enriching visual feature representation. For cross-modal alignment, CAT filters informative image-text pairs to reduce noise (Radenovic et al., 2023). Oscar (Li et al., 2020) improves image-text pairing by using object tags as anchor points, facilitating semantic alignment across modalities and achieving state-of-the-art results in vision-language tasks. Similarly, BLIP (Li et al., 2022a) removes noisy data by filtering image-text pairs based on similarity, followed by recaptioning to enhance data quality for downstream tasks.

In MMKG, few studies adapt visual integration to the unique characteristics of knowledge graphs, such as entity diversity, relationship complexity, and varying visual relevance across entities. Furthermore, knowledge graphs often contain entities that differ significantly in visual relevance, which means a one-size-fits-all approach to visual integration may introduce noise or dilute essential textual and structural information. This underscores the need for targeted methods like our DVMKG model, which adjusts visual integration based on entity-specific attributes to enhance knowledge representation.

B Prompts of LLM

In this study, we employed the prompt illustrated in Figure 7 to input key entity information, including Entity Name, Entity Description and the Entity Triple, into a large language model.

Entity Textual Descriptions: The first component includes the textual description of the entity in a template format, designed to convey the entity's defining characteristics. The template used for this purpose is: "[Entity]: [Definition Text]." This template standardizes the description input to the LLM, facilitating its interpretation of the entity's visualizability.

Entity Visualizability Score

Description:
Given the entity name, textual description, and its triple structure information from a multimodal knowledge graph, evaluate how easily the entity can be visually represented. Use the following criteria when assigning a visualization score:
NVE (Non-visualizable Entity): Entities that do not have a discernible visual form and cannot be represented visually, lacking a direct visual equivalent.
VE (Visualizable Entity): Entities that are easily visualized and have a distinct, concrete visual representation.

Scoring: 1. Assign a score ranging from 0 to 1, where 0 indicates a strong alignment with Non-visualizable Concepts and 1 indicates a strong alignment with Visualizable Concepts.
2. The score should reflect the degree to which an entity can be visually represented or conceptualized.

Input Format:
1. Entity Name: `{entity_name}`
2. Entity Description: `{entity_description}`
3. Entity Triple: `{triple_structure}`

Output Format:
Output should include the entity name followed by its visualizability score, formatted as:
1. Entity Name: `{entity_name}`
2. Score: `{score}`

Figure 7: Prompt for generating an Entity Visualizability Score.

Triple Structure Information: Each entity's contextual relationships within the knowledge graph are also considered, as these relationships often reveal the role and characteristics of the entity. To enhance the LLM's comprehension, we transform structured triples into natural language sentences using templates designed for each relation. For instance, the relation "current_team" is expressed in natural language as: "[X] plays for [Y]". This conversion enables the LLM to better interpret the entity's role within its relational context. For example, the triple "<Orlando_Engelaar current_team FC_Schalke_04>" is converted to: "Orlando Engelaar plays for FC Schalke 04."

C Experiments Setup

C.1 Datasets

To validate the performance of the proposed model, we set three widely public datasets: DB15K(Liu et al., 2019), YAGO15K(Liu et al., 2019) and FB15K-237(Mousselly-Sergieh et al., 2018), all

Datasets	#Ent.	#Rel.	#Train	#Valid	#Test
DB15K	12,842	279	79,222	9902	9904
YAGO15K	15,404	32	86,020	12,289	24,577
FB15K-237	14,541	237	272,115	17,535	20,466

Table 4: Datasets statistics.

of which contain three types of modal data related to their entities: structural information, images and text. Table 4 summarizes key dataset statistics, including entity and relation counts, and the sizes of the training, validation, and test sets.

Selection of the typical image (I_{typ}) in section 2.3.1. We prioritize using the entity's main cover image from its Wikipedia page. If such a cover image is not available for an entity (the entity is not on Wikipedia or lacks a cover image there), I_{typ} is then chosen through manual selection from the images associated with that entity within the dataset, picking the one deemed most visually representative.

C.2 Evaluation Indicators

To ensure consistency with previous studies and evaluate the performance of the DVMKG model, we select four common MKGC evaluation metrics: Mean Reciprocal Rank (MRR) and Hits@k (k = 1, 3, 10).

C.3 Baselines

To evaluate the DVMKG model, we compare it against SOTA methods from both KGC and MKGC domains. For unimodal methods, we select several key baselines, including TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ConvE (Dettmers et al., 2018), TuckER (Balažević et al., 2019), KC-GenRe (Wang et al., 2024) and MPIKGC (Xu et al., 2024a). For the multimodal methods, which serve as the main focus of comparison, we evaluate our model against several representative approaches, including IKRL (Xie et al., 2016), TransAE (Wang et al., 2019), RSME (Wang et al., 2021), VISTA (Lee et al., 2023), IMF (Li et al., 2023b), SNAG (Chen et al., 2024a) and AdaMF (Zhang et al., 2024b).

C.4 Implementation Details

All the experiments are run on a 125 G RAM computer with CPU Intel(R) Xeon(R) Silver 4110 CPU at 2.10 GHz and NVIDIA V100 GPU. All the reported results are the average of three different runs following the current practice. For the baselines in the main experiment, we ensured consistency by uniformly tuning hyperparameters, tuning the embedding dimension in {64, 128, 256, 512} and the number of negative samples in {16, 32, 64}. We optimize the model with Adam(Diederik, 2014) and the learning rate is tuned from $\{1e^{-3}, 1e^{-4}, 1e^{-5}\}$..

C.5 Multiple Modality Feature Encoding

We encode each modality in the MMKG to capture their unique characteristics, resulting in rich and complementary representations that serve as a robust foundation for subsequent fusion and learning processes.

Structural Modality. We employ the Long-CLIP (Zhang et al., 2024a) encoder to obtain the initial embeddings for both entities and relations based on their respective names. To fully exploit the graph's structural information, we enhance these embeddings with a Graph Attention Network (GAT) (Goodfellow et al., 2014), which aggregates information from neighboring nodes and relations, enriching them with higher-order dependencies and relational context.

Visual Modality. All images associated with each entity are encoded using the Long-CLIP encoder. Since entities may have multiple images, we aggregate these embeddings into a single visual representation per entity according to the RDS, ensuring that the selected images reflect both relevance and quality.

Textual Modality. We extract embeddings from textual descriptions using the Long-CLIP encoder, leveraging its ability to align visual and textual modalities within a shared embedding space.

D Effectiveness of Components in Representativeness Scoring Mechanisms.

To evaluate the effectiveness of each component in Equation 1 for calculating image representativeness scores, we defined a ranking task that assesses the extent to which the selected images represent the entity. In this task, for a given entity, the visual representation formed by its selected images was used as the query. The correct answer is the representation of the entity's name, which serves as the positive sample, while representations of 99 other randomly selected entity names were used as negative samples. The similarity between the entity's visual representation and each of the 100

Model	Representative Task		Diversity Task	
	Hits@1	MRR	Scene Diversity	Object Diversity
baseline	0.685	0.744	23.95	16.11
w/o S_{PVLM}	0.717	0.773	25.39	16.16
w/o S_{OD}	0.741	0.796	–	–
w/o S_{Un}	–	–	26.42	16.31
DVMKG	**0.774**	**0.825**	**27.06**	**16.35**

Table 5: Results for the representative image selection ranking task and diverse score task on YAGO15K.

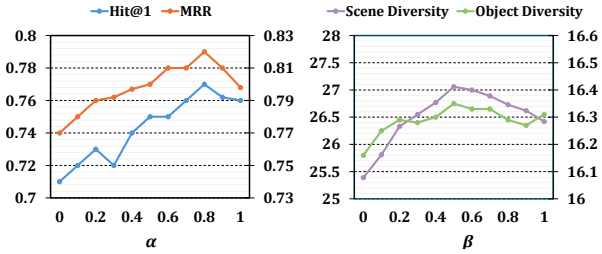

Figure 8: Hyperparameter for the representative image selection ranking task and diverse score task on YAGO15K.

entity representations (one positive, 99 negative) was computed, converting the problem into a ranking task. Key metrics include Hits@1 and MRR. The goal is to see how well the selected images allow the visual representation to rank the correct entity name representation higher than the negative samples.

Table 5 presents the performance of different methods on the representative image selection task. The S_{OD} represents the object detection in Equation 1. Compared to the baseline of averaging all available images (without selection), all selection methods show improved results. Object detection and PVLM individually enhance image selection quality, while the combined method achieves the highest scores on Hits@1 and MRR. The experimental results of the hyperparameter α combining Object Detection and PVLM are shown in Figure 8. By combining object detection and vision-language alignment, our method provides a more representative visual representation of entities.

E Effectiveness of Components in Diversity Scoring Mechanisms.

To evaluate the effectiveness of each component in calculating image diversity scores, we introduce two metrics: Object Diversity and Scene Diversity. This is derived based on scene distributions and KL divergence calculations.

Scene diversity is derived from scene classification probabilities for each image, obtained through a pretrained ResNet50 model (He et al., 2016) fine-tuned on the Places365 dataset (Zhou et al., 2017). Places365 provides a broad classification of 365 scene categories, including environments like bedrooms, restaurants, and beaches, which are used to represent diverse contexts. For each image, we compute a scene distribution $p(y|x)$, where $p(y|x)$ represents the probability of the image belonging to each scene category. This distribution is calculated by applying a softmax function to the raw output logits of ResNet50. To determine the diversity across all images associated with an entity, we calculate the average scene distribution $p(y)$ for the entity's image set. The diversity score is then quantified by computing the Kullback-Leibler (KL) divergence between the scene distribution of each image $p(y|x)$ and the average scene distribution $p(y)$. The final Diverse Score is obtained by taking the exponential of the expected KL divergence, expressed as:

$$IS = \exp\left(\mathbb{E}_x\left[KL(p(y|x)\|p(y))\right]\right). \quad (15)$$

Object diversity follows a similar procedure, except that instead of scene classification, we use an object detection model to identify objects within each image.

As shown in Table 5, the combined method $S_{Un} + S_{PVLM}$ outperforms all other methods in both Scene Diversity and Object Diversity. The experimental results of the hyperparameter β are shown in Figure 8. The synergy between the S_{Un} and S_{PVLM} methods effectively captures a wider range of scenes and objects. The experimental results confirm the effectiveness of the proposed diversity scoring mechanisms.