Are a Thousand Words Better Than a Single Picture? Beyond Images - A Framework for Multi-Modal Knowledge Graph Dataset Enrichment

Pengyu Zhang[1[0000−0001−5111−4487]], Klim Zaporojets[2[0000−0003−4988−978X]], Jie Liu[1[0000−0001−5116−5401]], Jia-Hong Huang[1,3[0000−0001−7943−2591]], and Paul Groth[1[0000−0003−0183−6910]]

[1] University of Amsterdam, Amsterdam, The Netherlands p.zhang@uva.nl
[2] Aarhus University, Aarhus, Denmark
[3] Amazon AGI, Seattle, USA

Abstract. Multi-Modal Knowledge Graphs (MMKGs) benefit from visual information, yet large-scale image collection is hard to curate and often excludes ambiguous but relevant visuals (e.g., logos, symbols, abstract scenes). We present **Beyond Images**, an automatic data-centric enrichment pipeline with optional human auditing. This pipeline operates in three stages: (1) large-scale retrieval of additional entity-related images, (2) conversion of all visual inputs into textual descriptions to ensure that ambiguous images contribute usable semantics rather than noise, and (3) fusion of multi-source descriptions using a large language model (LLM) to generate concise, entity-aligned summaries. These summaries replace or augment the text modality in standard MMKG models without changing their architectures or loss functions. Across three public MMKG datasets and multiple baseline models, we observe consistent gains (up to **+7%** Hits@1 overall). Furthermore, on a challenging subset of entities with visually ambiguous logos and symbols, converting images into text yields large improvements (**+201.35%** MRR and **+333.33%** Hits@1). Additionally, we release a lightweight Text-Image Consistency Check Interface for optional targeted audits, improving description quality and dataset reliability. Our results show that scaling image coverage and converting ambiguous visuals into text is a practical path to stronger MMKG completion. Code, datasets, and supplementary materials are available at `https://github.com/pengyu-zhang/Beyond-Images`.

Keywords: Multi-modal Knowledge Graphs · Link Prediction · Image-to-Text · Dataset Enrichment · Entity Representation.

1 Introduction

In Multi-Modal Knowledge Graphs (MMKGs), integrating text, images, audio, and video with structured triples yields richer signals and improves entity representations [5,12,24]. As illustrated on the left of Figure 1, the entity "`Amsterdam`"

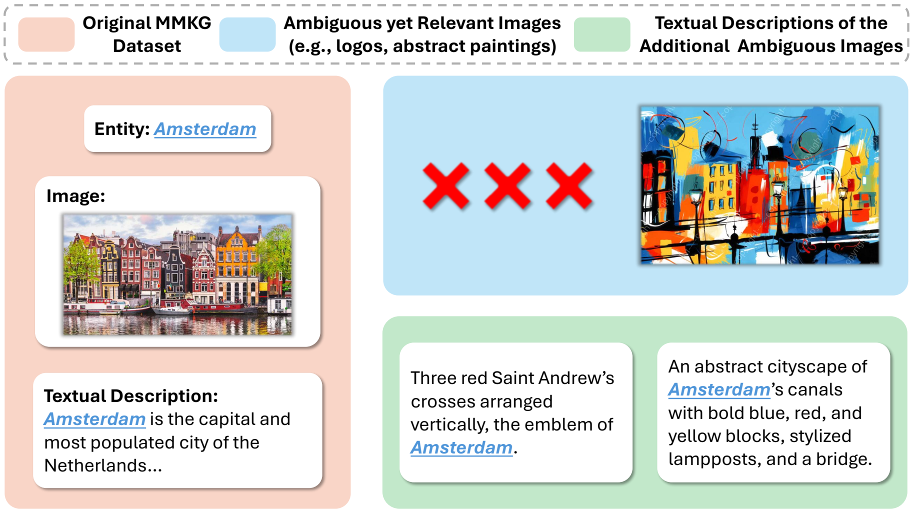

Original MMKG Dataset
Ambiguous yet Relevant Images (e.g., logos, abstract paintings)
Textual Descriptions of the Additional Ambiguous Images
Entity: Amsterdam
Image:
Textual Description: Amsterdam is the capital and most populated city of the Netherlands...
Three red Saint Andrew's crosses arranged vertically, the emblem of Amsterdam.
An abstract cityscape of Amsterdam's canals with bold blue, red, and yellow blocks, stylized lampposts, and a bridge.

Fig. 1. Original MMKG Dataset: the entity "`Amsterdam`" with its photo and a textual description. **Ambiguous yet Relevant Images**: additional visuals such as the three red Saint Andrew's crosses and an abstract cityscape, whose semantics may be unclear when incorporated directly as visual features. **Textual Descriptions of the Additional Ambiguous Images**: our framework converts these images into concise, entity-aligned text, expanding semantic coverage while mitigating noise from ambiguous visual embeddings.

may be associated with textual descriptions of its history and culture, photographs of canals and landmarks, and audio or visual content that reflects its urban atmosphere. Among these modalities, images often convey dense information, and when their content is unambiguous, downstream performance improves [22]. However, even though the Web enables rapid, large-scale collection of entity-related images, the construction process still suffers from two limitations.

Limitation 1: Lack of scalability. While most existing MMKG datasets automate the acquisition of textual descriptions, image collection often still relies on manual curation [31]. Curators or domain experts typically search for each entity and select a small set of images [20]. This practice helps avoid obviously misleading content but requires manual collection or verification and is difficult to scale to large graphs [17,2].

Limitation 2: Ambiguous yet relevant images remain unused. Manual curation typically favors highly representative images for each entity and excludes uncertain or ambiguous images [3,7]. Although this strategy supports quality control, it narrows the dataset's visual coverage and suppresses alternative yet valid information. The Web contains a large number of entity-related images whose visual semantics are inherently ambiguous. When incorporated directly into a dataset, such images often contribute noise rather than useful signal, which lowers data quality and degrades model performance [6,11]. Figure 1 highlights two common forms of image ambiguity: *Sparse-Semantic Images* (e.g., symbolic logos) are visually simple and lack sufficient semantic detail. Although sometimes domain-relevant, they rarely provide distinctive embeddings that benefit MMKG models [26,28]. In contrast, *Rich-Semantic Images* (e.g., abstract paintings) contain visually and semantically complex scenes that current embedding methods struggle to interpret, often leading to significant information loss [29]. Consequently, such images are frequently discarded during dataset construction. While this preserves highly reliable visual signals, it also reduces within-entity diversity. For example, in Figure 1, for the entity "`Amsterdam`", a dataset may only retain canal photographs but exclude the "triple red X" despite its historical and cultural significance to the city.

To address these limitations, we propose an automated framework, **Beyond Images**, which comprises three stages. First, the *Modality Extension Module* retrieves additional entity-related images beyond the original dataset, enabling large-scale expansion. Next, the *Semantic Generation Module* converts all retrieved images into textual descriptions, ensuring that ambiguous visuals contribute usable semantics rather than noise. Finally, the *LLM-based Semantic Fusion Module* consolidates multi-source descriptions, filters task-irrelevant content, and produces concise, entity-aligned summaries.

Our experiments show that our framework effectively addresses the two limitations defined above. In terms of scalability, we observe consistent improvements across three public MMKG datasets, achieving gains of up to +7% in Hits@1. To evaluate the impact of ambiguous yet relevant images (e.g., logo-only or symbol-centric visuals), we retain and convert newly retrieved images and measure link-prediction gains from their descriptions. On this challenging subset, performance increases by +201.35% MRR and +333.33% Hits@1. Additionally, we assess the quality of the generated image descriptions and show that our fused textual summaries outperform purely generated summaries in terms of overall quality. To support this evaluation, we implement a *Text-Image Consistency Check Interface* that allows for efficient targeted quality assessments by human annotators. Our main contributions are as follows:

(i) We propose **Beyond Images**, a reusable data-centric enrichment paradigm for MMKGs that converts ambiguous-yet-relevant visuals (e.g., logos or symbolic images) into entity-aligned textual representations and performs entity-level semantic fusion, without modifying downstream model architectures or training objectives.

(ii) We develop a lightweight *Text–Image Consistency Check Interface* to support optional human auditing, enabling low-effort quality inspection of generated entity summaries.

(iii) We release enriched datasets, code, and evaluation protocols, and conduct a systematic study across three datasets on four MMKG models, demonstrating consistent gains, up to +7% in Hits@1 overall, and +333.33% in Hits@1 on a challenging subset.

2 Related Work

Multi-Modal Knowledge Graphs. Link prediction is a core task in Knowledge Graphs (KGs), aiming to infer missing entities or relations from existing triples. Multi-Modal Knowledge Graphs (MMKGs) extend this setting by combining data from multiple modalities, such as text, images, and numerical features, to enhance representation learning and prediction accuracy. [20] introduced MMKGs that integrate numerical and visual information, demonstrating improvements in link prediction. Building on this, MCLEA [19] proposed a contrastive learning framework for multi-modal entity alignment. MCLEA first learns modality-specific representations and then applies contrastive learning to jointly model intra-modal and inter-modal interactions. Subsequently, MMKRL [21] incorporated a knowledge reconstruction module to integrate structured and multi-modal data into a unified space. This model also uses adversarial training to enhance robustness and performance. More recently, [14] introduced the MR-MKG method, which leverages MMKGs to improve reasoning capabilities in large language models. In parallel, [4] developed the SNAG model, which effectively combines structural, visual, and textual features, yielding improved link prediction performance. [35] leverages a multi-modal knowledge graph built from news text and detected visual objects to guide entity-aware image captioning.

Despite these advancements, MMKGs still face notable challenges. Many rely on manual data curation, which limits scalability and can introduce biases. Human experts tend to favor visually straightforward and highly representative images, potentially overlooking others that, although less explicit, could provide valuable additional information about the entity [23]. To address this gap, we propose an automated approach that retrieves and associates entity-related images from external sources without requiring manual annotation.

Automated Dataset Enrichment. Automated dataset enrichment is crucial in scaling the construction of MMKGs [10]. [1] introduced a method to align textual descriptions with knowledge graph entities, improving the semantic consistency of text embeddings. Building on this idea, [8] applied vision-language models and LLMs for visual question answering. Further advancements include the ADAGIO framework [30], which uses genetic programming to learn efficient augmentation frameworks for knowledge graphs, enhancing data augmentation processes. Similarly, [13] provides a lightweight automated knowledge graph construction solution by extracting keywords and evaluating relationships using graph Laplacian learning. Lastly, [25] further automated knowledge graph construction from unstructured text by integrating natural language processing techniques for entity extraction and relationship mapping, providing an end-to-end pipeline for converting raw text into structured knowledge. [11] proposed a multimodal deep-context knowledge extractor that integrates images via hierarchical captions and visual prefixes to enhance named entity recognition (NER) and relation extraction (RE) and build KGs.

However, most existing approaches still involve manual filtering or rely on domain experts to select images, making the process time-consuming and prone to subjective biases. To overcome these limitations, we introduce an automated framework that retrieves, filters, and converts images into textual representations. By transforming visual inputs into entity-aligned text and integrating them through semantic fusion, our approach streamlines dataset construction, reduces reliance on human intervention, and improves both scalability and consistency.

3 Methodology

This section presents our framework, **Beyond Images**. It consists of three key modules. The *Modality Extension Module* expands the dataset by automatically retrieving additional entity-related images from external sources. The *Semantic Generation Module* then converts both original and the newly retrieved images into textual descriptions to enrich semantic content. Finally, the *LLM-based Semantic Fusion Module* summarizes and filters these descriptions to reduce noise and retain task-relevant information. An overview of the full framework is shown in Figure 2.

Dataset structure (original vs. enriched). The original MMKG datasets provide: (i) a KG triple set with standard train/validation/test splits, (ii) human-written textual descriptions of the entities (used as the default text modality), and (iii) entity-associated images (raw images or image identifiers/URLs, depending on the dataset). To ensure consistent entity–image alignment, we standardize entity identifiers using Wikidata QIDs and normalize image filenames to the format `QID_index`. Based on the original data, our enriched dataset performs *entity-side attribute enrichment* without modifying the underlying triple structure. Specifically, we add: (i) automatically generated captions for the original images, (ii) automatically generated captions for newly retrieved images when the retrieval module is enabled, (iii) an entity-level fused summary produced by an LLM over all available captions for the entity (original + retrieved), and (iv) lightweight provenance metadata for retrieved images (e.g., source page URL, image URL, and extracted fields such as date/author when available). Downstream MMKG models do not consume the images or metadata directly; instead, they use the resulting textual fields as enhanced entity descriptions. Additional details on data processing and implementation are provided in the supplementary material.[4]

3.1 Modality Extension Module

To enhance the multi-modal coverage of MMKGs, we apply the *Modality Extension Module* to three public datasets: MKG-W,[5] MKG-Y,[6] and DB15K.[7] Each

[4] https://github.com/pengyu-zhang/Beyond-Images/tree/main/supplementary_material

[5] https://github.com/quqxui/MMRNS

[6] https://github.com/quqxui/MMRNS

[7] https://github.com/mniepert/mmkb

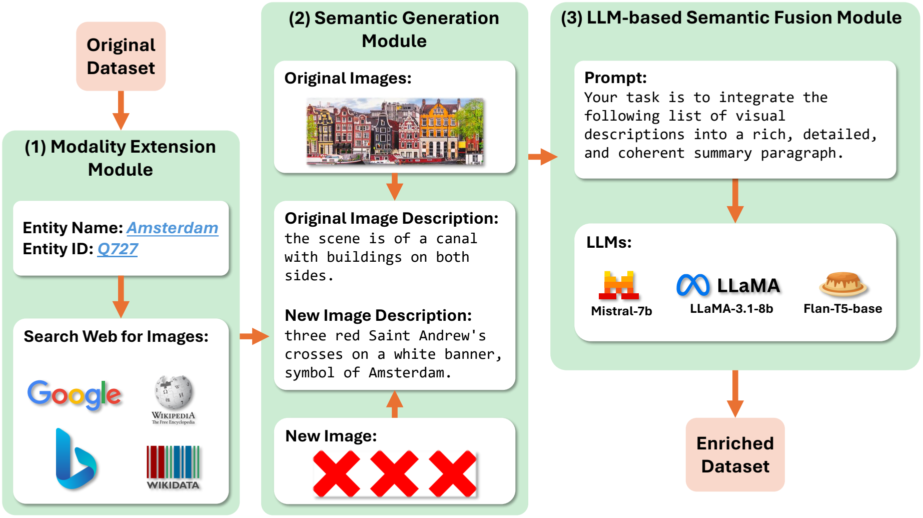

Original Dataset
(1) Modality Extension Module
Entity Name: Amsterdam
Entity ID: Q727
Search Web for Images:
Google
Wikipedia
Wikidata
(2) Semantic Generation Module
Original Images:
Original Image Description:
the scene is of a canal with buildings on both sides.
New Image Description:
three red Saint Andrew's crosses on a white banner, symbol of Amsterdam.
New Image:
(3) LLM-based Semantic Fusion Module
Prompt:
Your task is to integrate the following list of visual descriptions into a rich, detailed, and coherent summary paragraph.
LLMs:
LLaMA
Mistral-7b
LLaMA-3.1-8b
Flan-T5-base
Enriched Dataset

Fig. 2. Overview of our *Beyond Images* framework. Given an MMKG entity, (1) the *Modality Extension Module* retrieves additional web images from search engines; (2) the *Semantic Generation Module* produces per-image textual descriptions for both original and newly retrieved images; (3) the *LLM-based Semantic Fusion Module* consolidates valid descriptions into a single, rich summary paragraph via an explicit prompt. The summary becomes an enhanced textual view of the entity and is stored in the new enriched dataset for downstream MMKG completion.

dataset contains structured triples, entity descriptions, and images, covering diverse domains such as film, medicine, and geography. Our method is broadly compatible and applicable to any MMKG with visual content. Using entity names from each dataset, we crawled the corresponding English Wikipedia pages and collected all available images. For each image, we also retrieved metadata, including timestamps. Dataset statistics are summarized in Table 1. *Original Images* refer to those already included in the datasets, while *New Images* are retrieved by our framework. Across all datasets, the number of images per entity increases substantially, improving visual diversity and coverage.

Applying to other MMKGs. To run Beyond Images on a new MMKG dataset, we require only an entity list containing entity IDs and entity names (or canonical labels). The pipeline produces (i) per-image captions and (ii) an entity-level fused summary, which can be stored as an additional text attribute for each entity. Key configurable parameters include the number of retrieved images per entity, basic filtering rules (e.g., image size or type), and the choice of captioning model and fusion LLM. Once generated, the enriched textual fields can be reused across downstream models and tasks without additional LLM calls.

Table 1. Statistics of the three public MMKG datasets used in our experiments. *Original Images* are the image statistics provided by the source MMKGs. *New Images* are the additional images retrieved by our framework to extend coverage. "Avg Img" denotes the average number of images per entity; "Entity w/ Img" is the number of entities with at least one image.

	MKG-W	**MKG-Y**	**DB15K**
Entity	15,000	15,000	12,842
Relation	169	28	279
Train	34,196	21,310	79,222
Validation	4,276	2,665	9,902
Test	4,274	2,663	9,904
Text	14,123	12,305	9,078
Original Images			
Total Img	27,841	42,242	603,435
Avg Img	3.00	3.00	53.35
Entity w/ Img	9,285	14,099	11,311
New Images			
Total Img	81,323	56,646	176,858
Avg Img	5.81	4.23	14.58
Entity w/ Img	14,002	14,388	12,130

3.2 Semantic Generation Module

To capture semantic information from both original and newly collected multi-modal data, we introduce the *Semantic Generation Module*. This module uses three state-of-the-art image-to-text models: "*blip2-flan-t5-xxl*"[8] [15], "*git-large-coco*"[9] [27], and "*llava-v1.5-7b*"[10] [18] to generate rich textual descriptions from images. These models are selected for their strong generative capabilities and their ability to produce detailed, textual descriptions. In contrast, we exclude encoder-only models such as "*CLIP*"[11] from our framework, as they lack generative decoder and therefore cannot produce free-form text.

For each image, we generate a descriptive sentence using each of the three models and associate it with the corresponding entity in the dataset. To ensure simplicity and reproducibility, we use a single fixed prompt for all images: "Describe the scene, objects, colors, and other details in detail." We do not perform a prompt-level ablation in this paper; we discuss this as a limitation in Section 6.

3.3 LLM-based Semantic Fusion Module

While image-to-text models enrich multi-modal inputs with semantic details, they often introduce noise and task-irrelevant content. For instance, vision-language models can generate repetitive tokens (e.g., "person, person, person...") when processing certain images, or produce vague and redundant descriptions

[8] `https://huggingface.co/Salesforce/blip2-flan-t5-xxl`

[9] `https://huggingface.co/microsoft/git-large-coco`

[10] `https://huggingface.co/liuhaotian/llava-v1.5-7b`

[11] `https://github.com/openai/CLIP`

when the image is only loosely related to the target entity. This issue is particularly noticeable for entities associated with countries where the retrieved images are often maps or diagrams that provide limited value for entity representation learning.

To mitigate this issue, we introduce the *LLM-based Semantic Fusion Module*, which summarizes and aligns the visual descriptions with the semantics of the target entity in the MMKG. Specifically, we employ three LLMs: "*Flan-T5-base*",[12] "*LLaMA-3.1-8b-instruct*",[13] and "*Mistral-7b-instruct-v0.3*",[14] to summarize and filter the generated descriptions.

For each entity, we collect all image-based textual descriptions and feed them into the LLM using the following prompt: "Your task is to integrate the following list of visual descriptions for the entity '{entity_name}' into a rich, detailed, and coherent summary paragraph. Capture as many key details as possible, such as objects, colors, actions, and settings. Your final output must be a single paragraph, not a list." This process enables the LLM to extract informative features, remove redundancy, and produce a more coherent and entity-specific semantic summary. As a result, it improves the quality and relevance of the multi-modal inputs used for representation learning.

4 Experiments

This section presents our experimental setup and provides a comprehensive evaluation of the proposed framework on three widely used public MMKG datasets using four different model variants. Due to space constraints, implementation details are provided in the supplementary material.[15] Our evaluation focuses on the following five research questions (RQs):

RQ1 (Performance). Does the semantically enhanced dataset improve MMKG completion performance, particularly on ambiguous yet relevant image subsets? (Section 4.3)

RQ2 (Description Quality). What is the quality of the generated textual descriptions? (Section 4.4)

RQ3 (Ablation Studies). Can textual descriptions generated from images serve as an effective substitute for image embeddings? (Section 4.5)

RQ4 (Parameter Sensitivity). How do different image-to-text and LLM models influence performance, and how sensitive is the framework to these choices? (Section 4.6)

RQ5 (Case Study). Which types of triples benefit most from image-generated textual descriptions, and how does our framework affect these cases? (Section 4.7)

[12] https://huggingface.co/google/flan-t5-base

[13] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[14] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[15] https://github.com/pengyu-zhang/Beyond-Images/tree/main/supplementary_material

4.1 Evaluated Models

To ensure a thorough evaluation, we select four recent and widely used Multi-Modal Knowledge Graph (MMKG) models: **MMRNS**[16] [31], **MyGO**[17] [33], **NativE**[18] [32], and **AdaMF**[19] [34]. These models are chosen because they operate on datasets that include the original images, enabling a fair and consistent comparison. In addition, they are publicly available, widely recognized, and representative of recent progress in MMKG research. Rather than proposing a new model, our goal is to evaluate a *data-side* enrichment intervention. We therefore test the same enriched inputs across **four** MMKG architectures, and further vary the enrichment components by using **three** image-to-text captioning models and **three** fusion LLMs, all under the same training and evaluation protocols.

4.2 Task and Evaluation Setup

We evaluate link prediction on MMKGs. For each test triple, we construct two queries: *tail prediction* $(h, r, ?)$ and *head prediction* $(?, r, t)$. Given a query, the model scores all candidate entities $e \in \mathcal{E}$ and ranks them by the compatibility score $s(h, r, e)$ or $s(e, r, t)$.

The use of enriched descriptions. Our framework operates entirely on the data side. LLMs are used offline to convert images into text and to fuse multiple captions into a single entity-level summary. During training and inference, no LLM prompting is performed. The resulting summary is used as the textual modality input to standard MMKG models (e.g., MMRNS, MyGO, NativE, AdaMF) without changing their architectures or loss functions.

Metrics. We report Mean Reciprocal Rank (MRR) and Hits@K with $K \in \{1, 3, 10\}$, averaged over head and tail queries. MRR is the mean of 1/rank of the correct entity. Hits@K measures the fraction of queries where the correct entity appears in the top-K results. Higher values indicate better performance.

4.3 Overall Effectiveness (RQ1)

Overall Performance. We use the "*blip2-flan-t5-xxl*" model to generate textual descriptions from images and then apply "*Mistral-7b-instruct-v0.3*" to summarize all descriptions associated with each entity. This combination was selected due to its superior overall performance. A detailed analysis of how different image-to-text models and LLMs influence performance is provided in Section 4.5. The key link prediction results are reported in Table 2, which compares the performance of four MMKG models (MMRNS, MyGO, NativE, and AdaMF) across three datasets (MKG-W, MKG-Y, and DB15K) under different input settings.

In Table 2, the first row for each model reports results on the original dataset, reproduced from the corresponding papers. These results use only the default

[16] `https://github.com/quqxui/MMRNS`
[17] `https://github.com/zjukg/MyGO`
[18] `https://github.com/zjukg/NATIVE`
[19] `https://github.com/zjukg/AdaMF-MAT`

Table 2. Link prediction results on three MMKG datasets with four models (MMRNS, MyGO, NativE, AdaMF). Rows denote input settings: *Baseline* (original MMKG only), *G(o)* (textual descriptions from original images), *G(n)* (textual descriptions from newly retrieved images), *G(o+n)* (concatenation of *G(o)* and *G(n)*), and *Fusion* (an LLM summary generated from all image descriptions). Columns report MRR and Hits@K. Bold numbers indicate the best setting per model and dataset. The last row, *improv.(%)*, gives the relative gain of *Fusion* over the *Baseline* for each metric.

	MKG-W				**MKG-Y**				**DB15K**			
	MRR	**H@1**	**H@3**	**H@10**	**MRR**	**H@1**	**H@3**	**H@10**	**MRR**	**H@1**	**H@3**	**H@10**
MMRNS [31]												
Baseline	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47	32.68	23.01	37.86	51.01
G(o)	35.73	29.65	38.37	48.69	36.59	31.78	40.19	46.43	33.57	24.04	39.13	52.71
G(n)	36.13	29.93	38.58	49.02	36.93	31.96	40.33	46.58	33.37	23.78	39.02	52.40
G(o+n)	36.26	30.08	38.70	49.19	37.03	32.12	40.46	46.70	33.67	24.16	39.27	52.89
Fusion	**37.04**	**30.54**	**39.05**	**49.96**	**37.54**	**32.75**	**40.86**	**47.32**	**34.47**	**24.70**	**39.70**	**53.47**
improv.(%)	**+5.74**	**+6.82**	**+4.15**	**+5.24**	**+4.48**	**+7.26**	**+4.58**	**+4.06**	**+5.47**	**+7.33**	**+4.86**	**+4.82**
MyGO [33]												
Baseline	36.10	29.78	38.54	47.75	38.51	33.39	39.03	47.87	37.72	30.08	41.26	52.21
G(o)	37.19	30.85	39.65	48.75	39.63	34.73	39.88	48.90	38.84	31.53	42.37	53.74
G(n)	37.28	31.26	39.74	49.18	39.83	35.07	40.20	49.22	38.77	31.23	42.30	53.24
G(o+n)	37.42	31.42	39.88	49.35	39.97	35.26	40.32	49.37	38.97	31.69	42.49	53.92
Fusion	**38.05**	**31.82**	**40.54**	**49.82**	**40.77**	**35.69**	**40.78**	**50.01**	**39.38**	**32.24**	**43.06**	**54.22**
improv.(%)	**+5.41**	**+6.86**	**+5.19**	**+4.34**	**+5.87**	**+6.89**	**+4.47**	**+4.47**	**+4.40**	**+7.19**	**+4.36**	**+3.85**
NativE [32]												
Baseline	36.58	29.56	39.65	48.94	39.04	34.79	40.89	46.18	37.16	28.01	41.36	54.13
G(o)	37.37	30.56	40.44	49.93	39.63	35.95	41.93	47.03	38.68	28.83	42.48	55.11
G(n)	37.57	30.68	40.85	50.27	39.75	36.12	42.11	47.43	38.30	28.77	42.35	55.03
G(o+n)	37.69	30.80	40.97	50.41	39.83	36.27	42.25	47.56	38.84	28.92	42.61	55.22
Fusion	**38.04**	**31.43**	**41.42**	**51.04**	**40.38**	**36.92**	**42.72**	**48.30**	**39.55**	**29.37**	**43.12**	**55.62**
improv.(%)	**+3.98**	**+6.33**	**+4.45**	**+4.30**	**+3.43**	**+6.13**	**+4.47**	**+4.59**	**+6.42**	**+4.84**	**+4.27**	**+2.76**
AdaMF [34]												
Baseline	35.85	29.04	39.01	48.42	38.57	34.34	40.59	45.76	35.14	25.30	41.11	52.92
G(o)	36.92	30.16	39.78	49.34	39.79	35.37	41.45	46.41	36.20	26.24	42.29	54.35
G(n)	37.20	30.35	39.77	49.73	40.05	35.86	41.89	46.78	35.85	26.08	42.13	54.24
G(o+n)	37.36	30.50	39.85	49.88	40.21	36.04	42.02	46.88	36.32	26.34	42.43	54.51
Fusion	**38.04**	**31.29**	**40.39**	**50.46**	**40.54**	**36.68**	**42.36**	**47.48**	**36.74**	**26.98**	**42.89**	**54.84**
improv.(%)	**+6.11**	**+7.76**	**+3.54**	**+4.21**	**+5.10**	**+6.82**	**+4.37**	**+3.77**	**+4.56**	**+6.63**	**+4.33**	**+3.63**

textual descriptions and image embeddings provided in the original datasets. The remaining rows evaluate four configurations, each adding a single additional input to the original dataset: "*G(o)*", "*G(n)*", "*G(o+n)*", or "*Fusion*". "*G(o)*" uses textual descriptions generated from the original images. "*G(n)*" uses textual descriptions generated from newly downloaded images. "*G(o+n)*" combines both sources by concatenating their descriptions. "*Fusion*" uses an LLM to summarize all descriptions from both original and new images into a single paragraph, which is then used as input. "*H@K*" refers to Hits at K, and "*improv.(%)*" indicates the relative performance gain, computed as: Boost $= \frac{\text{Fusion Result}-\text{Baseline Result}}{\text{Baseline Result}}$.

The results in Table 2 show that using the enriched datasets consistently improves performance across all metrics (MRR, Hits@1, Hits@3, and Hits@10) for every model. For example, MyGO achieves 5.41% and 6.86% improvements in MRR and Hits@1 respectively on the MKG-W dataset. Similar improvements

Table 3. Performance on a subset of entities whose image sets *include logos or symbols*. *Baseline* uses the original MMKG without converting these images. *Fusion* applies our framework to convert the logo-like images into textual descriptions. Results show that our framework recovers usable semantics, yielding large gains in MRR and Hits@K.

	MRR	**H@1**	**H@3**	**H@10**
Baseline	13.89	7.50	15.00	27.50
Fusion	**41.87**	**32.50**	**47.50**	**57.50**
improv.(%)	**+201.35**	**+333.33**	**+216.67**	**+109.09**

are observed on MKG-Y and DB15K, indicating that the proposed approach generalizes across different models and datasets. These results confirm the benefits of incorporating image-based textual descriptions into MMKG tasks.

Among all configurations, "*Fusion*" consistently achieves the best performance. Compared with the simple concatenation strategy in "*G(o+n)*", the LLM-generated summary provides additional gains, suggesting that naive aggregation fails to filter out noise and redundancy. In contrast, summarizing the image descriptions using an LLM produces more coherent and semantically aligned content with respect to the entity, leading to improved downstream performance.

It is worth noting that while "*Fusion*" performs best on all datasets, the comparison between "*G(o)*" and "*G(n)*" shows a dataset-specific pattern. On MKG-W and MKG-Y, "*G(n)*" outperforms "*G(o)*", whereas on DB15K the opposite trend is observed. We attribute this to the fact that DB15K contains more original images than newly downloaded ones (see Table 1). As a result, the model benefits from the larger number of descriptions in "*G(o)*", leading to stronger entity representations and higher performance.

Ambiguous yet Relevant Image Subsets. We manually sampled 20 entities whose images consist of logos or other abstract marks and formed a small evaluation subset as shown in Table 3. We compared two inputs: the "*Baseline*", which uses only the original MMKG, and "*Fusion*", which converts these logo images into textual descriptions and fuses them with the existing modalities. Although logos provide limited visual information, converting them into text supplies the missing semantics. This yields large gains for link prediction, improving MRR and Hits@K by substantial margins (e.g., +201.35% MRR, +333.33% Hits@1).

Training Efficiency. We analyze the training efficiency of the proposed approach. As shown in the supplementary material,[20] once generated, the enriched data can be reused across models and tasks without extra cost. Training time increases by only 7-30 minutes, while performance improves, demonstrating a favorable cost-benefit trade-off.

[20] `https://github.com/pengyu-zhang/Beyond-Images/tree/main/supplementary_material`

4.4 Description Quality (RQ2)

Our framework performs enrichment automatically in an end-to-end pipeline (retrieval → captioning → fusion). The following step is optional and serves only to audit the accuracy of the generated summaries. Inspired by the CoT Curation Toolkit,[21] we release a lightweight browser interface for human verification (demo and code[22]). The interface shows, side by side, the LLM summary and the full image set for a given entity. Annotators select a verdict (Match, Mismatch, or Uncertain), add a brief rationale, and may hide irrelevant images.

From each dataset, we draw a random sample of 100 cases for manual auditing. A case is correct when the summary captures the main visual semantics of the image set; we do not require the summary to identify or verify the entity itself. Under this criterion, we observe no clear mismatches and *two cases* where the description is inaccurate or incomplete. These results indicate that the generated summaries reliably reflect the visual content of the images across datasets.

4.5 Modality Contribution (RQ3)

Figure 3 reports Hits@1 on the MKG-W dataset for four models (MMRNS, MyGO, NativE, AdaMF) under three different modality combinations (see further). Bars represent Hits@1 (higher is better). The x-axis denotes the modality combinations: I+T uses original image embeddings and text; T+G uses text plus image-generated descriptions; I+T+G uses all three. The legend indicates the source of the image-generated descriptions used in each bar: *G(o)* from original images, *G(n)* from newly retrieved images, *G(o+n)* as their concatenation, and *LLM Fusion* as an LLM summary over all descriptions.

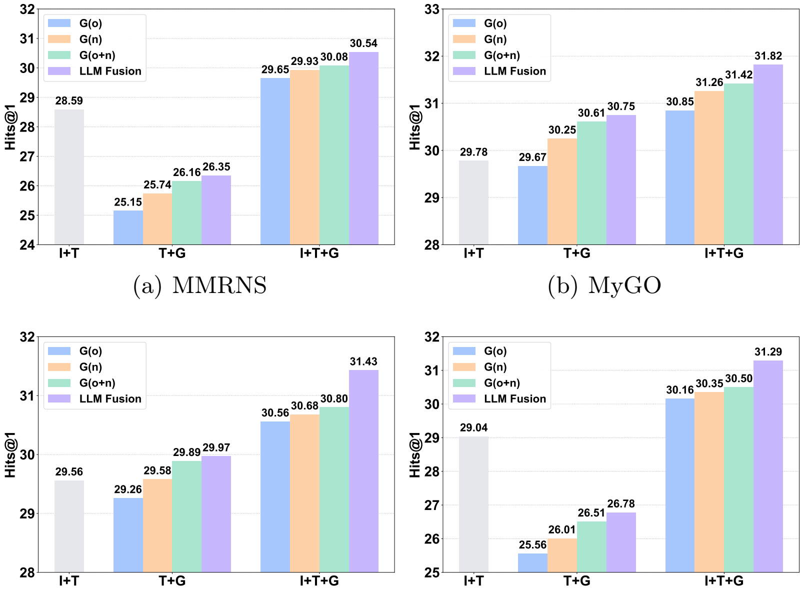
As shown in the figure, using only two modalities results in limited model performance. Performance improves when all three modalities are combined, with the best results observed in the I+T+G with *LLM Fusion*. This demonstrates the benefit of integrating complementary information across modalities, which helps close the semantic gap and improves prediction accuracy.

We observe different levels of reliance on visual embeddings versus image generated text across models. For MyGO and NativE, replacing image embeddings with generated descriptions (T+G vs. I+T) yields comparable or higher Hits@1. In contrast, MMRNS and AdaMF show performance degradation under T+G relative to I+T, indicating a stronger dependence on visual features. A plausible explanation is architectural: models like MMRNS and AdaMF use global image embeddings that preserve high-level semantics, while models like MyGO and NativE tokenize images into local patches, which can dilute entity-level semantics, making textual descriptions a competitive substitute.

Many MMKG images are *ambiguous yet relevant*, which provides a complementary, data-side explanation for the model-dependent behaviors observed

[21] https://github.com/caocongfeng/CoT_curation_toolkit

[22] https://github.com/pengyu-zhang/Beyond-Images/tree/main/video_demo



(a) MMRNS

(b) MyGO

(c) NativE

(d) AdaMF

Fig. 3. Hits@1 comparison on the MKG-W dataset across four models under three modality settings: *I+T* (image embeddings + textual descriptions), *T+G* (textual descriptions + image-generated descriptions), and *I+T+G* (all). Legend: *G(o)* denotes descriptions from original images, *G(n)* from newly retrieved images, *G(o+n)* their concatenation, and *LLM Fusion* an LLM summary over all descriptions. Bars report Hits@1 (higher is better).

above. For *sparse-semantic* visuals such as logos and symbols, raw visual embeddings can be weak or non-discriminative, making it difficult for a completion model to extract relation-relevant cues. For *rich-semantic* but abstract or stylistic images, fixed visual encoders may miss key contextual details, leading to information loss. In both cases, converting images into text helps make the latent semantics more explicit and easier to align with KG relations, particularly when multiple captions are fused at the entity level to reduce redundancy and noise. Motivated by this intuition, we next conduct a controlled modality ablation to better isolate how much each modality contributes under a fixed architecture.

To isolate the contribution of each modality, and inspired by prior work [36,16], we evaluate the MyGO model on the MKG-W dataset under three settings (as shown in Table 4): *Image Only*, *Text Only* (the original MMKG entity textual descriptions, *not* our image-derived captions), and *Image + Text*. *Image Only* attains the lowest scores (e.g., MRR 31.70; H@1 24.47), indicating that images alone provide limited relational signal for link prediction. *Text Only* improves performance across all metrics (MRR 35.93; H@1 28.41), suggesting that textual descriptions encode stronger cues about entities and relations. *Image +*

Table 4. Modality ablation on the MKG-W dataset. We compare *Image Only*, *Text Only*, and *Image + Text*. Text alone outperforms image alone across all metrics, while fusing both modalities yields the best MRR and Hits@K.

	MRR	**H@1**	**H@3**	**H@10**
Image Only	31.70	24.47	32.34	41.68
Text Only	35.93	28.41	36.75	45.09
Image + Text	**36.10**	**29.78**	**38.54**	**47.75**

Table 5. Performance comparison of different pre-trained image-to-text generation models on the MKG-W dataset using the MyGO model. "***H@K***" stands for "Hits at K."

	MRR	**H@1**	**H@3**	**H@10**
Image-to-Text Models (no LLM Fusion)				
Baseline	36.10	29.78	38.54	47.75
Git-large-coco	36.59	30.61	38.57	48.02
Llava-v1.5-7b	36.21	30.15	38.40	47.83
Blip2-flan-t5-xxl	**37.42**	**31.42**	**39.88**	**49.35**
LLM Fusion (using Blip2 for image-to-text)				
Baseline	36.10	29.78	38.54	47.75
Flan-T5-base	37.58	31.48	39.98	49.42
LLaMA-3.1-8b-instruct	37.87	31.56	40.02	49.61
Mistral-7b-instruct-v0.3	**38.05**	**31.82**	**40.54**	**49.82**

Text achieves the best performance (MRR 36.10; H@1 29.78), confirming that the two modalities are complementary and that fusing them yields consistent gains. Although the improvements over *Text Only* are moderate, they highlight the incremental yet useful contribution of visual evidence.

4.6 Impact of Model Choices (RQ4)

In this section, we analyze how the choice of image-to-text models and LLMs affects the overall performance of our framework. We first evaluate the impact of different pre-trained image-to-text models by applying them to the MyGO baseline on the MKG-W dataset. As shown in the upper block of Table 5, all models improve performance compared to the original baseline, with "*blip2-flan-t5-xxl*" achieving the highest gains across all metrics.

Next, we study the effect of different LLMs used to summarize the generated image descriptions. Using "*blip2-flan-t5-xxl*" as the image-to-text backbone, we compare three LLMs: "*Flan-T5-base*", "*LLaMA-3.1-8b-instruct*", and "*Mistral-7b-instruct-v0.3*". As shown in the lower block of Table 5, all three models outperform the baseline, with "*Mistral-7b-instruct-v0.3*" achieving the best results. These findings demonstrate the effectiveness of LLM-based summarization in refining image-derived descriptions and enhancing semantic alignment.

To illustrate these differences, we present a representative example: the album cover of "`Hot Sauce Committee Part Two`" by the "`Beastie Boys`". The "*blip2-flan-t5-xxl*" model generates brief yet informative descriptions such as "The cover of beastboys hot sauce committee part two" and "The logo for beastie boys is shown in black and white." These outputs capture both the type of image (album cover) and key named entities (artist and album title), which helps the model align visual and textual information more effectively. This alignment improves entity disambiguation and link prediction.

```
blip2-flan-t5-xxl: ‘‘The cover of Beastie Boys Hot Sauce Committee Part Two’’.
```

The "*git-large-coco*" model only outputs a generic phrase, failing to identify any specific entities. While it correctly recognizes the image type, it lacks the detail needed for fine-grained semantic alignment.

```
git-large-coco: ‘‘cover art for the album’’.
```

In comparison, "*llava-v1.5-7b*" generates a long and visually rich description, focusing on colors, shapes, and layout.

```
llava-v1.5-7b: ‘‘The image is a colorful collage featuring a variety of shapes, colors,
and patterns. The collage is composed of multiple squares, each containing different
colors and designs. The squares are arranged in a visually appealing manner, creating
a vibrant and dynamic composition. The colors in the collage are diverse, with some
squares featuring red, green, blue, yellow, and orange hues. The patterns and shapes
within the squares vary, adding depth and interest to the overall design. The collage
appears to be a creative and artistic representation of the artist’s vision, showcasing
their talent and skill in combining different elements to create a cohesive and visually
engaging piece’’.
```

The different outputs of the three models are likely due to differences in architecture, language modeling, and training objectives. The superior performance of "*blip2-flan-t5-xxl*" suggests that generating straightforward and semantically aligned descriptions is beneficial for representation learning and downstream tasks.

4.7 Case Analysis (RQ5)

To demonstrate how image-generated descriptions enhance model performance, we compare predictions made with the original dataset inputs against those made with our enhanced inputs that additionally include image-generated textual descriptions. The largest gain is observed for the triple *(Hot Sauce Committee Part Two, performer, Beastie Boys)*. Here, the head entity is *Hot Sauce Committee Part Two*, the relation is *performer*, and the tail entity is *Beastie Boys*. With the relation *performer* and the textual description of the tail entity *Beastie Boys*, the rank of the correct head entity improves from 13,680 to 1,330. Likewise, when given the head entity *Hot Sauce Committee Part Two*, its textual description, and the relation *performer*, the rank of the correct tail entity improves from 11,435 to 4,628. This case illustrates that adding image-derived text can strengthen entity alignment by making relation-relevant cues more explicit.

In the original dataset, the associated visuals for this triple are sparse-semantic (e.g., logo-like or symbolic content). Visual embeddings therefore tend to capture abstract shapes and patterns with limited entity-discriminative information, making it difficult for the model to learn meaningful connections from images alone. In contrast, the generated textual descriptions translate such ambiguous visuals into explicit cues (e.g., names or identity hints), which can be more directly aligned with KG relations and thus improve ranking. Due to copyright and reuse restrictions, we do not reproduce the corresponding images in the paper. The entities can be inspected via their public knowledge base pages (e.g., Wikidata[23] [24]).

5 Conclusion and Future Work

We introduced **Beyond Images**, an automatic enrichment pipeline for MMKGs that scales image retrieval, converts visuals into task-aligned text, and fuses multi-source descriptions into concise, entity-aligned summaries. This *data-side* approach is model-agnostic and can be integrated with standard MMKG models without modifying their architectures or loss functions. Experiments on three public datasets show consistent improvements (up to +**7%** Hits@1 overall). On a challenging subset with logo or symbol images, converting images into text substantially boosts performance (+**201.35%** MRR; +**333.33%** Hits@1), indicating that images with weak visual signals can still convey strong semantics when rendered as text. We additionally provide a lightweight *Text-Image Consistency Check Interface* for optional, low-effort human auditing to further enhance dataset reliability.

For future work, we plan to (i) integrate temporal signals (e.g., timestamps and versioned descriptions) to study entity evolution [9], (ii) explore stronger and multilingual captioning and fusion LLMs to extend beyond English-centric corpora, (iii) incorporate active sampling for more efficient human auditing, and (iv) extend the framework to additional modalities (e.g., audio and video) and open-vocabulary entities. We hope that our released code, datasets, and auditing interfaces facilitate reproducible research on scalable, data-centric MMKG enrichment.

6 Limitations

Our enrichment quality depends on the reliability of off-the-shelf image-to-text models and fusion LLMs, which may produce incomplete or noisy descriptions for challenging images. We currently generate a single sentence per image and use a single fixed prompt, without systematically studying alternative prompting or multi-aspect captioning strategies. Finally, while we provide an optional human-auditing interface, our main experiments focus on link prediction; broader downstream applications (e.g., QA or entity-centric retrieval) are left for future work.

[23] https://en.wikipedia.org/wiki/File:Hot_Sauce_Committee_Part_Two.png

[24] https://commons.wikimedia.org/wiki/File:Beastie_Boys_logo_(1985-1986).png

Acknowledgments. The first author is supported by the China Scholarship Council (NO. 202206540007) and the University of Amsterdam. This funding source had no influence on the study design, data collection, analysis, or manuscript preparation and approval. This work was partially supported by the EU's Horizon Europe programme within the ENEXA project (grant Agreement no. 101070305). This project has received funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101146515 KG-PROVENANCE.

Disclosure of Interests. The authors declare that they have no competing interests relevant to the content of this article.

References

1. An, B., Chen, B., Han, X., Sun, L.: Accurate text-enhanced knowledge graph representation learning. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 745–755. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). `https://doi.org/10.18653/v1/N18-1068`
2. Chen, H., Shen, X., Lv, Q., Wang, J., Ni, X., Ye, J.: SAC-KG: Exploiting large language models as skilled automatic constructors for domain knowledge graph. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4345–4360. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). `https://doi.org/10.18653/v1/2024.acl-long.238`
3. Chen, J., Gao, Y., Ge, M., Li, M.: Ambiguity-aware and high-order relation learning for multi-grained image–text matching. Knowledge-Based Systems **316**, 113355 (2025)
4. Chen, Z., Fang, Y., Zhang, Y., Guo, L., Chen, J., Pan, J.Z., Chen, H., Zhang, W.: Noise-powered multi-modal knowledge graph representation framework. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp. 141–155. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025)
5. Chen, Z., Zhang, Y., Fang, Y., Geng, Y., Guo, L., Chen, X., Li, Q., Zhang, W., Chen, J., Zhu, Y., Li, J., Liu, X., Pan, J.Z., Zhang, N., Chen, H.: Knowledge graphs meet multi-modal learning: A comprehensive survey (2024)
6. Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z.: Deep learning based synthesis of mri, ct and pet: Review and analysis. Medical image analysis **92**, 103046 (2024)
7. Firmansyah, A.F., Zahera, H.M., Sherif, M.A., Moussallem, D., Ngomo, A.C.N.: Ants: Abstractive entity summarization in knowledge graphs. p. 133–151. Springer-Verlag, Berlin, Heidelberg (2025). `https://doi.org/10.1007/978-3-031-94575-5_8`, `https://doi.org/10.1007/978-3-031-94575-5_8`
8. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B.A., Tao, D., Hoi, S.C.H.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10867–10877 (2022)

9. Huang, S., Poursafaei, F., Danovitch, J., Fey, M., Hu, W., Rossi, E., Leskovec, J., Bronstein, M., Rabusseau, G., Rabbany, R.: Temporal graph benchmark for machine learning on temporal graphs. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 2056–2073. Curran Associates, Inc. (2023)
10. Klironomos, A., Zhou, B., Zheng, Z., Mohamed, G.E., Paulheim, H., Kharlamov, E.: Realite: Enrichment of relation embeddings in knowledge graphs using numeric literals. In: The Semantic Web: 22nd European Semantic Web Conference, ESWC 2025, Portoroz, Slovenia, June 1–5, 2025, Proceedings, Part I. p. 41–58. Springer-Verlag, Berlin, Heidelberg (2025). https://doi.org/10.1007/978-3-031-94575-5_3, https://doi.org/10.1007/978-3-031-94575-5_3
11. Ko, H., Yoo, J., Jeong, O.R.: Mdcke: Multimodal deep-context knowledge extractor that integrates contextual information. Alexandria Engineering Journal **119**, 478–492 (2025). https://doi.org/https://doi.org/10.1016/j.aej.2025.01.119, https://www.sciencedirect.com/science/article/pii/S1110016825001474
12. Koloski, B., Pollak, S., Navigli, R., Škrlj, B.: Automl-guided fusion of entity and llm-based representations for document classification. In: Pedreschi, D., Monreale, A., Guidotti, R., Pellungrini, R., Naretto, F. (eds.) Discovery Science. pp. 101–115. Springer Nature Switzerland, Cham (2025)
13. Kuo, C.W., Kira, Z.: Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 17948–17958 (2022)
14. Lee, J., Wang, Y., Li, J., Zhang, M.: Multimodal reasoning with multimodal knowledge graph. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10767–10782. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.579
15. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (23–29 Jul 2023)
16. Li, X., Zhao, X., Xu, J., Zhang, Y., Xing, C.: Imf: interactive multimodal fusion model for link prediction. In: Proceedings of the ACM web conference 2023. pp. 2572–2580 (2023)
17. Li, Y., Tian, Y., Huang, Y., Lu, W., Wang, S., Lin, W., Rocha, A.: Fakescope: Large multimodal expert model for transparent ai-generated image forensics. arXiv preprint arXiv:2503.24267 (2025)
18. Lin, B., Tang, Z., Ye, Y., Huang, J., Zhang, J., Pang, Y., Jin, P., Ning, M., Luo, J., Yuan, L.: Moe-llava: Mixture of experts for large vision-language models (2024)
19. Lin, Z., Zhang, Z., Wang, M., Shi, Y., Wu, X., Zheng, Y.: Multi-modal contrastive representation learning for entity alignment. In: Calzolari, N., Huang, C.R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.S., Ryu, P.M., Chen, H.H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T.K., Santus, E., Bond, F., Na, S.H. (eds.) Proceedings of the 29th International Conference on Computational Linguistics. pp. 2572–2584. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022)
20. Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., Rosenblum, D.S.: Mmkg: Multi-modal knowledge graphs. In: The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings. p.

459–474. Springer-Verlag, Berlin, Heidelberg (2019). https://doi.org/10.1007/978-3-030-21348-0_30

21. Lu, X., Wang, L., Jiang, Z., He, S., Liu, S.: Mmkrl: A robust embedding approach for multi-modal knowledge graph representation learning. Applied Intelligence **52**(7), 7480–7497 (May 2022). https://doi.org/10.1007/s10489-021-02693-9
22. Mei, K., Talebi, H., Ardakani, M., Patel, V.M., Milanfar, P., Delbracio, M.: The power of context: How multimodality improves image super-resolution. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23141–23152 (2025)
23. Misra, I., Zitnick, C.L., Mitchell, M., Girshick, R.: Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels . In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2930–2939. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.320
24. Purohit, D., Chudasama, Y., Rivas, A., Vidal, M.E.: Sparkle: Symbolic capturing of knowledge for knowledge graph enrichment with learning. In: Proceedings of the 12th Knowledge Capture Conference 2023. p. 44–52. K-CAP '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3587259.3627547
25. Rezayi, S., Zhao, H., Kim, S., Rossi, R., Lipka, N., Li, S.: Edge: Enriching knowledge graph embeddings with external text. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2767–2776. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.221
26. Su, T., Zhang, X., Sheng, J., Zhang, Z., Liu, T.: Loginmea: Local-to-global interaction network for multi-modal entity alignment. In: ECAI 2024, pp. 1173–1180. IOS Press (2024)
27. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language (2022)
28. Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., Wang, H., Jiang, S.: Logo-2k+: A large-scale logo dataset for scalable logo classification. Proceedings of the AAAI Conference on Artificial Intelligence **34**(04), 6194–6201 (Apr 2020). https://doi.org/10.1609/aaai.v34i04.6085
29. Wilber, M.J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., Belongie, S.: Bam! the behance artistic media dataset for recognition beyond photography. In: Proceedings of the IEEE international conference on computer vision. pp. 1202–1211 (2017)
30. Xiang, Y., Zhang, Z., Chen, J., Chen, X., Lin, Z., Zheng, Y.: OntoEA: Ontology-guided entity alignment via joint knowledge graph embedding. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1117–1128. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.96
31. Xu, D., Xu, T., Wu, S., Zhou, J., Chen, E.: Relation-enhanced negative sampling for multimodal knowledge graph completion. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 3857–3866. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548388
32. Zhang, Y., Chen, Z., Guo, L., Xu, Y., Hu, B., Liu, Z., Zhang, W., Chen, H.: Native: Multi-modal knowledge graph completion in the wild. In: Proceedings of the 47th

International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 91–101. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3657800

33. Zhang, Y., Chen, Z., Guo, L., Xu, Y., Hu, B., Liu, Z., Zhang, W., Chen, H.: Tokenization, fusion, and augmentation: towards fine-grained multi-modal entity representation. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press (2025). https://doi.org/10.1609/aaai.v39i12.33454, https://doi.org/10.1609/aaai.v39i12.33454
34. Zhang, Y., Chen, Z., Liang, L., Chen, H., Zhang, W.: Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 17120–17130. ELRA and ICCL, Torino, Italia (May 2024)
35. Zhao, W., Wu, X.: Boosting entity-aware image captioning with multi-modal knowledge graph. Trans. Multi. **26**, 2659–2670 (Jan 2024). https://doi.org/10.1109/TMM.2023.3301279, https://doi.org/10.1109/TMM.2023.3301279
36. Zhao, Y., Cai, X., Wu, Y., Zhang, H., Zhang, Y., Zhao, G., Jiang, N.: MoSE: Modality split and ensemble for multimodal knowledge graph completion. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 10527–10536. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.719, https://aclanthology.org/2022.emnlp-main.719/