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

Anonymous ACL submission

Abstract

Multi-Modal Knowledge Graphs (MMKGs) enhance entity representations by incorporating text, images, audio, and video, offering a more comprehensive understanding of each entity. Among these modalities, images are especially valuable due to their rich content and the ease of large-scale collection. However, many images are semantically unclear, making it challenging for the models to effectively use them to enhance entity representations. To address this, we present the ***Beyond Images*** framework, which generates textual descriptions for entity images to more effectively capture their semantic relevance to the associated entity. By adding textual descriptions, we achieve up to 5% improvement in Hits@1 for link prediction task across three MMKG datasets. Furthermore, our scalable framework reduces the need for manual construction by automatically extending three MMKG datasets with additional images and their descriptions. Our work highlights the importance of textual descriptions for MMKGs. Our code and enriched datasets are publicly available at `https://anonymous.4open.science/r/Beyond-Images-2266`.

1 Introduction

Knowledge Graphs (KGs) organize information in a structured format, where entities are represented as nodes and relationships as edges (Ma, 2022). This structure enables efficient data storage and reveals connections across diverse knowledge (Wan et al., 2024). Multi-Modal Knowledge Graphs (MMKGs) incorporate additional modalities, e.g., images, audio, and video, to enrich entity representations (Chen et al., 2024b). As illustrated in Figure 1, an entity like "***Amsterdam***" can be described through text about its history and culture, images showcasing its canals and landmarks, and audio or video capturing its atmosphere. By combining multiple modalities, MMKGs provide richer and more distinctive entity representations, thereby improving downstream tasks such as link prediction and KG completion (Koloski et al., 2025).

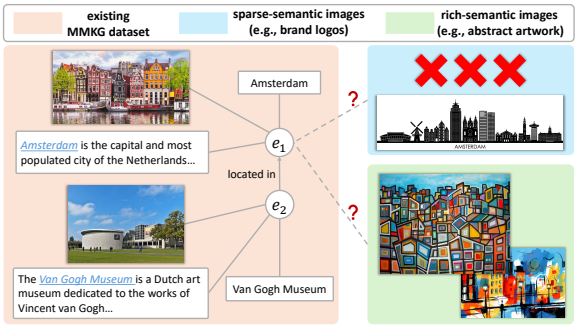


Figure 1: An illustration of a Multi-Modal Knowledge Graph (MMKG), where entities like "***Amsterdam***" and "***Van Gogh Museum***" can be described through multiple modalities (e.g., text, images). Some images (such as **logos** or **abstract artwork**) may be semantically unclear, causing ambiguous image embeddings in MMKGs.

Images are important in MMKGs due to their rich content and ease of large-scale collection. However, most existing MMKG datasets and models suffer from challenges in capturing the full semantics of images in relation to the entities they represent. Concretely, global feature extraction methods (e.g., Convolutional Neural Networks) generate high-level representations for entire images (Li et al., 2022), while local feature extraction methods (e.g., Vision Transformers) divide images into patches for more fine-grained processing (Dosovitskiy et al., 2020). Both approaches face challenges with two types of images: ***sparse-semantic images*** (e.g., brand logos) offer limited distinguishing features (Su et al., 2024; Wang et al., 2020), while ***rich-semantic images*** (e.g., abstract artwork) contain complex semantics that are difficult to capture accurately (Wilber et al., 2017) (see Appendix A).

To address this challenge, we study whether textual descriptions can serve as a more effective alternative by asking: ***Are a thousand words better than a single picture?*** This is especially important

when images lack clear semantic relevance to entities. Our automated framework standardizes and enriches MMKG datasets by converting both existing and newly collected images into textual descriptions. This text-driven approach reduces the impact of semantically ambiguous images by adding meaningful textual descriptions to the model. Additionally, MMKG dataset construction is ***highly labor-intensive***, as it often requires domain experts to filter and validate images manually, making the process time-consuming and easily influenced by individual biases (Chen et al., 2024a). Our framework addresses this by supporting large-scale image collection while reducing the need for labor-intensive dataset creation and manual filtering. In some models, replacing images with their textual descriptions results in more efficient and compact representations, highlighting the importance of textual descriptions in MMKGs. Our contributions are as follows:

- We generate meaningful textual descriptions from images to preserve semantic information. This text-driven approach highlights the role of textual descriptions in MMKGs.
- Experiments on three datasets with four models, which show a 2%–5% improvement in Hits@1 for link prediction, confirming that text-based image representations enhance model accuracy.
- A framework automating large-scale image collection, reducing reliance on experts and eliminating individual biases in dataset construction.

2 Related Work

2.1 Multi-Modal Knowledge Graphs

Multi-Modal Knowledge Graphs (MMKGs) combine data from multiple modalities, such as text, images, and numerical features, to enhance tasks such as link prediction and knowledge completion. Liu et al. (2019) introduced MMKGs that use numerical and visual information, demonstrating improvements in link prediction. Building on this, Lin et al. (2022) introduced the MCLEA model, which leverages contrastive learning to integrate multi-modal information for entity alignment. MCLEA first learns modality-specific representations and then applies contrastive learning to jointly model intra-modal and inter-modal interactions. Further progress was made with the MMKRL model (Lu et al., 2022), which includes 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, (Lee et al., 2024) introduced the MR-MKG method, which uses MMKGs to improve reasoning capabilities in large language models. Additionally, (Chen et al., 2025) developed the SNAG model, which effectively combines structural, visual, and textual features, leading to better results in knowledge graph link prediction.

Despite these advancements, MMKGs still face notable challenges. Many rely on manual data curation, which limits scalability and can introduce biases. Human experts often prefer straightforward images, potentially overlooking others that, while less obvious, could provide valuable additional information about the entity (Misra et al., 2016). To address this gap, we propose an automated approach that associates images with entities without manual intervention.

2.2 Automated Dataset Enrichment

Automated dataset enrichment is crucial in scaling the construction of MMKGs. An et al. (2018) introduced a method that connects textual descriptions with knowledge graph entities, improving the semantic consistency of text embeddings. Building on this, Guo et al. (2022) proposed an approach that uses pre-trained vision-language models to generate textual descriptions from images, enriching multi-modal datasets effectively. Further advancements include the ADAGIO framework (Xiang et al., 2021), which uses genetic programming to learn efficient augmentation frameworks for knowledge graphs, enhancing data augmentation processes. Similarly, Kuo and Kira (2022) provides a lightweight automated knowledge graph construction solution by extracting keywords and evaluating relationships using graph Laplacian learning. Lastly, Rezayi et al. (2021) automates knowledge graph creation from unstructured text by integrating natural language processing techniques for entity extraction and relationship mapping, providing an end-to-end solution for converting raw text into structured knowledge.

However, most of these methods still involve some degree of manual filtering or rely on domain experts to choose images, making the process time-consuming and eliminating individual biases. To overcome these limitations, we introduce a fully

	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
Img w/ Timestamp	0	0	0
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
Img w/ Timestamp	55,317	39,281	124,721
Entity w/ Img	14,002	14,388	12,130

Table 1: Overview of three public MMKG datasets, summarizing key statistics including the number of entities, relations, dataset splits, and image attributes. The table details original images, newly downloaded images, average images per entity, and images with timestamps.

automated framework that retrieves, filters and converts images into textual representations. This approach simplifies dataset construction while reducing reliance on human intervention, improving both scalability and consistency.

3 Methodology

This section introduces our framework for automatically enriching three widely used public Multi-Modal Knowledge Graph (MMKG) datasets with extracted images and their corresponding textual descriptions. The framework is shown in Figure 2. Additionally, we provide statistics on the original and updated datasets to illustrate the enrichment achieved through our framework in Table 1.

3.1 Datasets Information

We used three widely adopted public datasets to validate our approach: MKG-W,[1] MKG-Y,[2] and DB15K.[3] Each dataset contains three key components: structured knowledge, textual descriptions, and images. Details of the dataset are provided in Table 1.

MKG-W (Multi-modal KG-Wikipedia) is a dataset constructed by extracting structured knowledge from Wikipedia (Sun et al., 2020). Textual descriptions were obtained from DBpedia and aligned with the corresponding entities using additional `sameAs` links provided in the same work. Images were further extended using web search engines and manually screened by human experts (Xu et al., 2022). The dataset contains 15,000 samples, each identified by its corresponding Wikidata URL.

MKG-Y (Multi-modal KG-YAGO) is a dataset constructed by extracting structured knowledge from YAGO (Sun et al., 2020). Similar to MKG-W, textual descriptions were obtained from DBpedia and aligned with entities using `sameAs` links. Images were extended using web search engines and manually screened by experts (Xu et al., 2022). The dataset consists of 15,000 samples, each identified by its corresponding YAGO entity name.

DB15K (MMKB-DB15K) is an open-source MMKG introduced in (Liu et al., 2019). Its structured knowledge is derived from a subset of DBpedia, as described in (Lehmann et al., 2015). Since the original dataset lacked textual descriptions for entities, (Xu et al., 2022) extended it with textual information from DBpedia. Images were collected using search queries based on entity names, notable types, and Wikipedia URIs. This image acquisition process involved multiple steps. First, images were retrieved from three search engines: Google Images, Bing Images, and Yahoo Images, storing up to 20 top-ranked results per entity. Second, images smaller than 224 pixels or with an extreme aspect ratios (one side 2.5 times larger than the other) were filtered out. Third, corrupted, low-quality, and duplicate images (determined by pixel-wise similarity below a predefined threshold) were removed. Finally, the remaining images images were scaled to a maximum height or width of 500 pixels while maintaining their aspect ratio. The DB15K dataset contains 12,842 samples, with each sample named according to its corresponding DBpedia URL.

3.2 Standardizing and Aligning Original Images

Many existing datasets provide only image embeddings rather than raw images, and those that do include images use inconsistent naming conventions, making alignment challenging. To address this, we standardized image naming by mapping all entities to their corresponding Wikidata IDs (QIDs). This ensures compatibility across systems and facilitates dataset expansion. Additionally, this approach also enables the retrieval of additional images and metadata from Wikipedia for further enrichment. Details are in Table 1 and Appendix B.

[1] https://github.com/quqxui/MMRNS
[2] https://github.com/quqxui/MMRNS
[3] https://github.com/mniepert/mmkb

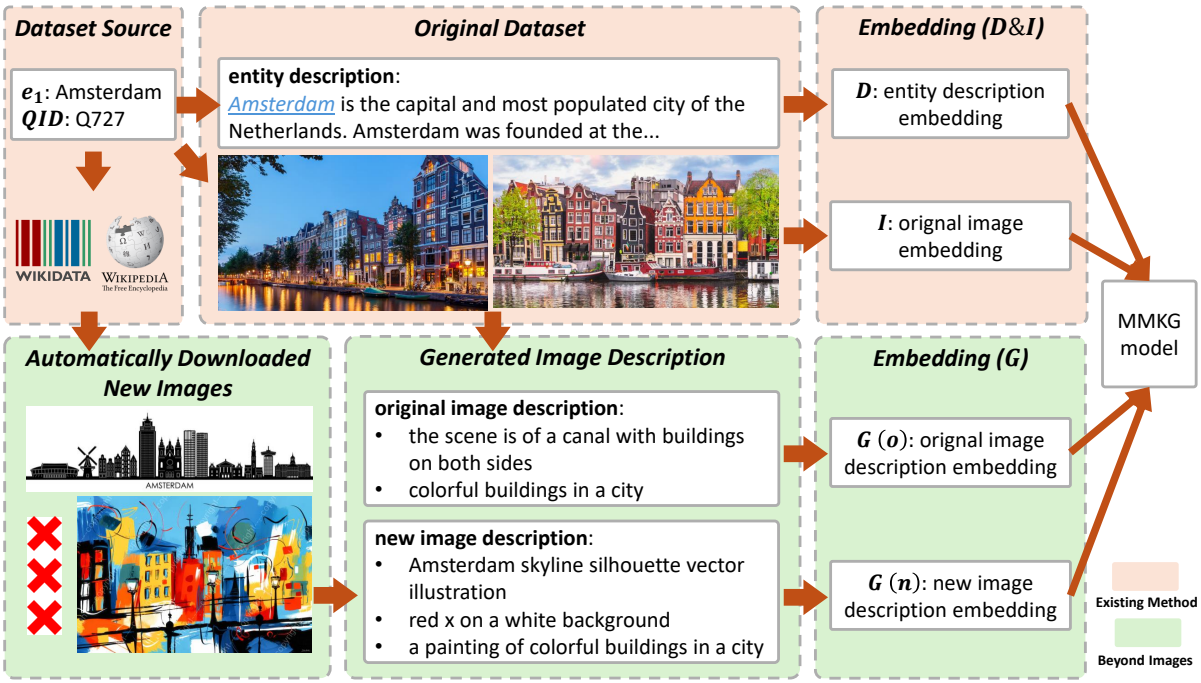


Figure 2: Overview of our framework. Each entity in the dataset is linked to its corresponding Wikidata ID (***Dataset Source***), automatically downloads additional images from Wikipedia (***Automatically Downloaded New Images***), and uses pre-trained model BLIP-2 to generate textual descriptions for both original and newly downloaded images (***Generated Image Description***). These descriptions are then embedded to produce ***G(o)*** and ***G(n)***, which, along with entity description embeddings ***D*** and image embeddings ***I***, are used as inputs to the MMKG model.

3.3 Downloading New Images

With the obtained Wikidata URLs for all three datasets, we accessed the corresponding English Wikipedia pages for each sample and retrieved all associated images. For each image, we accessed its details page to download the file along with relevant metadata (e.g., timestamps). Details of newly downloaded images are provided in Table 1.

3.4 Generating Textual Descriptions for Original and New Images

We used the BLIP-2[4] (Li et al., 2023) model to generate textual descriptions from images. BLIP-2 efficiently bridges images and text by integrating a frozen image encoder with a frozen language model, connected through a lightweight Querying Transformer. This design is particularly well-suited for our task, as it effectively translates visual features into meaningful textual descriptions while minimizing computational overhead and eliminating the need for extensive retraining.

In our implementation, we used the "***blip2-flan-t5-xxl***" model to generate textual descriptions from images. To efficiently process images, we implemented a function called "***generate_batch_descriptions***", which handles input images in batches and generates detailed textual descriptions for each image. This function takes as input a list of image paths and a prompt (prompt="Describe the scene, objects, colors, and other details in detail".) to guide the model in generating comprehensive semantic descriptions.

During processing, images in each batch were encoded as tensors using a processor and passed through the model to generate textual descriptions. These descriptions and their corresponding filenames were saved to an output file, providing meaningful textual data for MMKG tasks. The final output files are summarized in Appendix C.

The dataset provides detailed information for each image, such as URLs, metadata, and automatically generated textual descriptions. Each entry consists of a unique identifier (`id`), links to the corresponding Wikipedia page (`page_url`) and image file (`image_url`), metadata extracted from the image (`table_data`), and textual descriptions generated by BLIP-2 (`image_blip2_detail`). The metadata includes various attributes such as date (`Formatted_Date`), author, and resolution, which are crucial for MMKG research.

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

3.5 Using Both Original and Enriched Datasets as Model Inputs

Existing MMKG models typically use three types of inputs: structured knowledge, textual descriptions, and images. In this work, we extended the input by incorporating two additional types of textual descriptions: one generated from the dataset's original images and another generated from newly downloaded images (Section 3.4).

This modification affected only model inputs, leaving the loss function unchanged. This simple adjustment improved model performance, demonstrating the effectiveness of incorporating textual descriptions from images.

4 Experiments

This section introduces our experimental setup and presents a detailed discussion of the experiments conducted to evaluate our framework on three widely used public MMKG datasets using four different models. Implementation details can be found in Appendix D. In our experiments, we primarily focus on the following two Research Questions (RQs):

RQ1. Is a thousand words better than a single picture? Does using ***Beyond Images*** to standardize and enrich datasets lead to improved model performance? (Section 4.2)

RQ2. Can textual descriptions generated from images effectively replace image embeddings? Specifically, in which cases do text-based representations serve as a suitable alternative to image embeddings? (Section 4.3)

4.1 Evaluated Models

To provide a comprehensive evaluation, we employ four Multi-Modal Knowledge Graph (MMKG) models: MMRNS, MyGO, NativE, and AdaMF. These models were chosen because they use datasets containing the original images we were able to retrieve, thereby ensuring fairness and consistency.

MMRNS[5] (Xu et al., 2022) enhances MMKG completion through a knowledge-guided cross-modal attention mechanism and contrastive semantic sampling. By integrating relational embeddings, it improves the representation of both positive and negative samples, leading to significant performance improvements on multimodal KG benchmarks.

MyGO[6] (Zhang et al., 2024b) introduces fine-grained tokenization and contrastive learning technique to improve multi-modal entity representations. Its cross-modal entity encoder effectively captures complex interactions among modalities, at the time of publication this method surpassed 19 recent models.

NativE[7] (Zhang et al., 2024a) addresses imbalanced modality distributions by employing a dual adaptive fusion module combined with modality adversarial training. It achieved state-of-the-art results across diverse datasets while ensuring efficiency and generalizability.

AdaMF[8] (Zhang et al., 2024c) employs adaptive modality weights and modality-adversarial training to tackle modality imbalance in MMKGs. It achieves superior multi-modal fusion and outperforms 19 recent methods, establishing new state-of-the-art results on MMKGC benchmarks.

4.2 Main Results (RQ1)

The main results are shown in Table 2, which summarizes the link prediction performance of four models (MMRNS, MyGO, NativE, and AdaMF) across three datasets (MKG-W, MKG-Y, and DB15K) under different settings.

Model performance is evaluated using rank-based metrics, including Mean Reciprocal Rank (MRR) and Hits@K (K = 1, 3, 10). MRR calculates the average of the reciprocal ranks of the correct answers in the predicted ranking list, while Hits@K measures the proportion of correct answers appearing within the top K predictions. Both metrics are commonly used in evaluating link prediction tasks, with higher scores indicating better model performance. Full results are in Appendix E.

In Table 2, the first row for each model presents the experimental results on the original datasets, as reproduced from the original papers. "***D***" represents entity descriptions, "***I***" denotes image embeddings, "***G(o)***" refers to textual descriptions generated from original images, and "***G(n)***" corresponds to textual descriptions from newly downloaded images using our ***Beyond Images*** framework. "***Improvement (↑%)***" represents the percentage increase (Boost $= \frac{\text{Our Result} - \text{Baseline Result}}{\text{Baseline Result}}$) in performance of the enriched datasets compared to the original datasets.

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

[6] https://github.com/zjukg/MyGO

[7] https://github.com/zjukg/NATIVE

[8] https://github.com/zjukg/AdaMF-MAT

Model	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	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47	32.68	23.01	37.86	51.01
MMRNS (D+I+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
MMRNS (D+I+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
MMRNS (D+I+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**
Improvement (↑%)	**3.50%**	**5.20%**	**3.21%**	**3.62%**	**3.07%**	**5.21%**	**3.56%**	**2.71%**	**3.04%**	**5.01%**	**3.71%**	**3.69%**
MyGO	36.10	29.78	38.54	47.75	38.51	33.39	39.03	47.87	37.72	30.08	41.26	52.21
MyGO (D+I+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
MyGO (D+I+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
MyGO (D+I+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**
Improvement (↑%)	**3.66%**	**5.51%**	**3.48%**	**3.35%**	**3.80%**	**5.60%**	**3.31%**	**3.13%**	**3.31%**	**5.36%**	**2.99%**	**3.27%**
NativE	36.58	29.56	39.65	48.94	39.04	34.79	40.89	46.18	37.16	28.01	41.36	54.13
NativE (D+I+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
NativE (D+I+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
NativE (D+I+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**
Improvement (↑%)	**3.02%**	**4.21%**	**3.32%**	**3.01%**	**2.01%**	**4.25%**	**3.32%**	**2.98%**	**4.52%**	**3.25%**	**3.02%**	**2.02%**
AdaMF	35.85	29.04	39.01	48.42	38.57	34.34	40.59	45.76	35.14	25.30	41.11	52.92
AdaMF (D+I+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
AdaMF (D+I+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
AdaMF (D+I+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**
Improvement (↑%)	**4.21%**	**5.02%**	**2.15%**	**3.02%**	**4.25%**	**4.95%**	**3.52%**	**2.46%**	**3.35%**	**4.12%**	**3.20%**	**3.00%**

Table 2: Link prediction results of four models across three datasets. "***D***" represents entity descriptions, "***I***" denotes image embeddings, "***G(o)***" refers to textual descriptions generated from original images, and "***G(n)***" corresponds to textual descriptions from newly downloaded images. "***H@n***" stands for "Hits at n." The "***Improvement (↑%)***" indicates the performance gain of the best-performing model (highlighted in bold) over the baseline model.

Table 2 demonstrates that using the enriched datasets improves performance across all metrics (MRR, Hits@1, Hits@3, and Hits@10) for every model. For example, the MyGO model achieves a 3.66% and 5.51% boost in MRR and Hits@1. Similar trends are observed on the MKG-Y and DB15K datasets, further confirming the general applicability and effectiveness of our method across different datasets and models. These results highlight the importance of incorporating textual descriptions generated from images to enhance MMKG tasks.

It is important to note that, as shown in Table 2, the four models exhibit better improvements on the DB15K dataset when using textual descriptions generated from the original images provided in the dataset. This result differs from the trends observed for the MKG-W and MKG-Y datasets. We hypothesize this difference arises because the DB15K dataset contains more original images than the number of images automatically downloaded using our ***Beyond Images*** framework (as detailed in Table 1). Hence, for the DB15K dataset, the performance of "***G(o)***" surpasses that of "***G(n)***", because the larger number of generated textual descriptions in "***G(o)***", compared to "***G(n)***", enables the model to learn richer and more accurate entity representations.

4.3 Ablation Study (RQ2)

Figure 3 shows the performance differences of four models on the MKG-W dataset when using various combinations of input modalities. The y-axis represents Hits@1, while the x-axis corresponds to the following input scenarios. "***D***": Textual descriptions provided for entities in the original dataset. "***I***": Image embeddings of the entities in the original dataset. "***G***": T,extual descriptions generated for images by the ***Beyond Images*** framework. In the "***G***" scenario, "***o***" denotes descriptions generated from original images, while "***n***" indicates descriptions generated from newly downloaded images.

The figure shows that model performance is relatively poor when only two modalities are used. However, performance improves when all three modalities are combined, with the best results achieved when "***D+I+G(o+n)***" are used together. This highlights the benefits of multimodal integration, which enables models to fully leverage complementary information across modalities, effectively mitigating potential semantic gaps in individual modalities and achieving superior prediction results.

Additionally, we observe varying impacts of textual and visual information across different models. In MyGO and NativE, textual information has a

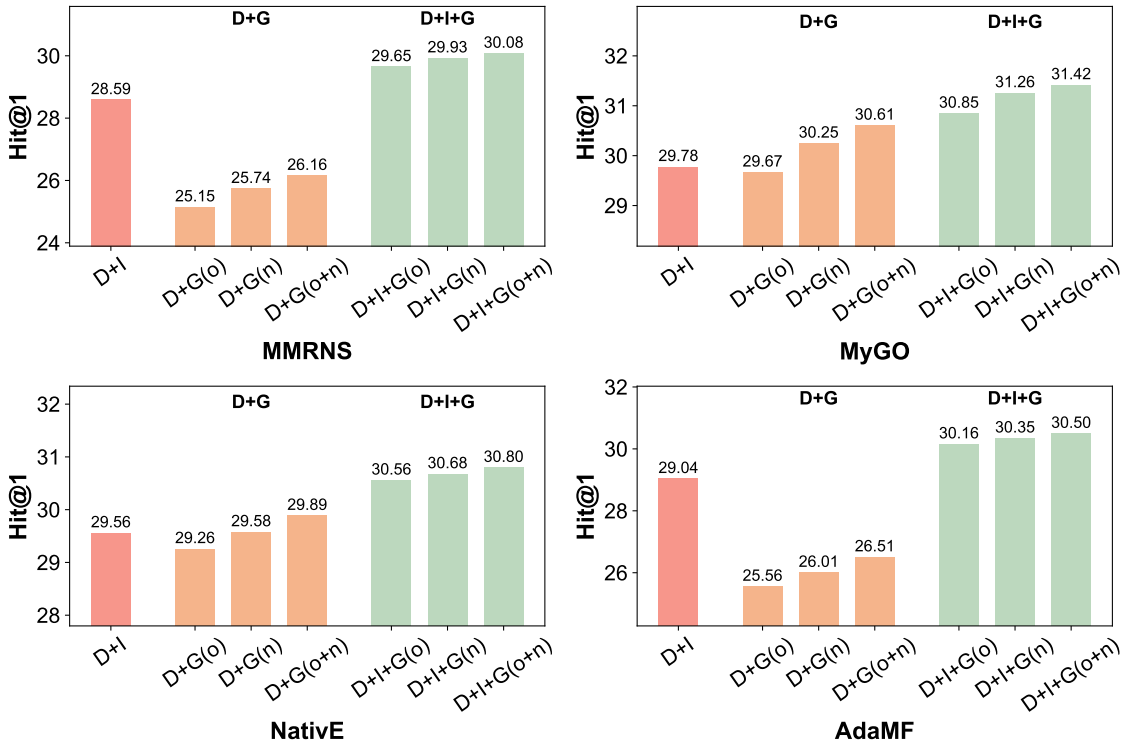


Figure 3: Hits@1 comparison across four models on the MKG-W dataset under different modality combinations. "***D***" represents entity descriptions, "***I***" denotes image embeddings, "***G(o)***" refers to textual descriptions from original images, and "***G(n)***" corresponds to those from newly downloaded images. The importance of textual and visual information varies across models. Models using global image embeddings (MMRNS, AdaMF) benefit more from visual inputs, while those using tokenized image embeddings (MyGO, NativE) rely more on textual descriptions.

more significant impact than visual information (textual information > visual information). In contrast, for MMRNS and AdaMF, visual information plays a more prominent role (visual information > textual information). We believe that this difference is driven by how images are processed and integrated into the representation space within each model. In MMRNS and AdaMF model, global image embeddings encode an entire image into a single vector, preserving high-level semantic information that aligns well with entity-level reasoning in knowledge graphs. This approach ensures that key visual attributes relevant to the entity are retained, often making visual information more influential than textual descriptions. In contrast, MyGO and NativE split image into multiple discrete visual tokens, shifting from a global to a localized perspective. While this allows for fine-grained feature extraction, it may disrupt semantic coherence at the entity level, reducing the effectiveness of visual information for knowledge graph reasoning. Additionally, tokenized embeddings may introduce redundancy or irrelevant details.

Each approach has advantages depending on the task in MMKGs. Global image embeddings work well for entity recognition and concept alignment, while tokenized embeddings are better for fine-grained visual reasoning, such as object interactions and spatial relationships. The impact of textual and visual information varies with the representation method: global embeddings enhance visual contributions, whereas textual descriptions become more important in detailed multi-modal tasks. These findings emphasize the importance of selecting the appropriate modality representation strategy for specific applications.

In these ablation experiments, single-modality (i.e., only "***D***", "***G***" or "***I***") tests were not conducted, as MMKG completion relies on the synergy of multiple modalities. A single-modality alone cannot provide comprehensive knowledge completion. The goal is to evaluate the contribution of each modality rather than test their standalone effectiveness. By comparing models with and without specific modalities, we can more accurately assess the benefits of multimodality.

4.4 Case Analysis: Boosting Performance with Textual Descriptions

To demonstrate how generated textual descriptions enhance model performance, we compared the predictions of the "***D+I***" and "***D+I+G***" configurations. The most significant improvement was observed in the triple "***(Hot Sauce Committee Part Two, performer, Beastie Boys)***". In this example, the head entity is "***Hot Sauce Committee Part Two***", the relation is "***performer***", and the tail entity is "***Beastie Boys***". When the model was given the relation "***performer***" along with the textual description of the tail entity "***Beastie Boys***", the correct head entity's rank improved significantly from 13,680 to 1,330. Likewise, when provided with the head entity "***Hot Sauce Committee Part Two***", its textual description, and the relation "***performer***", the rank of the correct tail entity improved from 11,435 to 4,628. This demonstrates the effectiveness of adding textual descriptions generated from images, as they enhance entity representation alignment.

(a) ***Hot Sauce Committee Part Two*** image 1: "The cover of beastboys hot sauce committee part two".

(b) ***Hot Sauce Committee Part Two*** image 2: "The scene shows a group of men walking on a bridge".

(c) ***Beastie Boys*** image 1: "three men are leaning on a stair railing".

(d) ***Beastie Boys*** image 2: "The logo for beastie boys is shown in black and white".

Figure 4: Triple: ***(Hot Sauce Committee Part Two, performer, Beastie Boys)***. Images (a) and (b) correspond to the head entity ***Hot Sauce Committee Part Two***, while images (c) and (d) represent the tail entity ***Beastie Boys***. The textual descriptions generated by BLIP-2 capture key semantic details like the entity name and album title. This helps the model better align entity representations, significantly improving performance.

Figure 4 presents the existing images and the textual descriptions generated for this triple. These images highlight the challenge of "sparse-semantic images (e.g., brand logos)," where visual embeddings primarily capture abstract shapes and patterns, providing limited semantic information. Such images often lack distinctive features, making it difficult for the model to learn meaningful connections. As a result, relying solely on image embeddings may lead to weak or inaccurate entity representations. However, incorporating textual descriptions generated from these images helps mitigate this limitation. Specifically, the generated text "***G***" in this example includes details such as the entity name and album title, allowing the model to better align the entity with other modalities ("***D***" and "***I***").

5 Conclusion and Future Work

In this paper, we introduced ***Beyond Images***, an automated framework for standardizing and enriching Multi-Modal Knowledge Graph (MMKG) datasets by generating meaningful textual descriptions for both existing and newly collected images. This text-driven approach enhances semantic representations across a spectrum of images, from those with sparse-semantics (e.g., brand logos) to those with rich-semantics (e.g., abstract artwork), mitigating the impact of semantic unclarity. Additionally, it removes the need for extensive expert filtering, making dataset construction more scalable.

We evaluated ***Beyond Images*** using four state-of-the-art models on three public datasets, achieving 2%-5% improvement in the link prediction task. These results highlight the importance of language in MMKGs, especially when images are weakly related to entities. Furthermore, the proposed textual representations derived from image descriptions offer a more compact and efficient alternative to images for some of the evaluated models.

For future work, we plan to extend this framework to additional datasets to support large-scale MMKG research and further validate our approach. Additionally, we aim to investigate more complex tasks that incorporate temporal information (Huang et al., 2023). Concretely, we plan to explore the impact of image timestamps on dynamic entity evolution, to gain deeper insights into knowledge evolution in multi-modal environments.

Limitations

While our ***Beyond Images*** framework improves Multi-Modal Knowledge Graphs (MMKGs) by converting images into textual descriptions, it has certain limitations.

First, the quality of the generated textual descriptions relies on the pre-trained vision-language model (BLIP-2). If the model produces inaccurate descriptions, the semantic alignment between images and entities may be weakened, potentially impacting downstream tasks.

Second, while replacing images with text reduces dependence on semantically ambiguous images, it may also lead to the loss of fine-grained visual details. These details could be important for specific tasks, such as visual reasoning in MMKGs.

References

Bo An, Bo Chen, Xianpei Han, and Le Sun. 2018. Accurate text-enhanced knowledge graph representation learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 745–755, New Orleans, Louisiana. Association for Computational Linguistics.

Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. 2024a. SAC-KG: Exploiting large language models as skilled automatic constructors for domain knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4345–4360, Bangkok, Thailand. Association for Computational Linguistics.

Zhuo Chen, Yin Fang, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen, and Wen Zhang. 2025. Noise-powered multi-modal knowledge graph representation framework. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 141–155, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Albert Li, Dacheng Tao, and Steven C. H. Hoi. 2022. 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)*, pages 10867–10877.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. 2023. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems*, volume 36, pages 2056–2073. Curran Associates, Inc.

Boshko Koloski, Senja Pollak, Roberto Navigli, and Blaž Škrlj. 2025. Automl-guided fusion of entity and llm-based representations for document classification. In *Discovery Science*, pages 101–115, Cham. Springer Nature Switzerland.

Chia-Wen Kuo and Zsolt Kira. 2022. 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)*, pages 17948–17958.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10767–10782, Bangkok, Thailand. Association for Computational Linguistics.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2022. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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*, page 459–474, Berlin, Heidelberg. Springer-Verlag.

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*, 52(7):7480–7497.

Xiaogang Ma. 2022. Knowledge graph construction and application in geosciences: A review. *Computers & Geosciences*, 161:105082.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2939, Los Alamitos, CA, USA. IEEE Computer Society.

Saed Rezayi, Handong Zhao, Sungchul Kim, Ryan Rossi, Nedim Lipka, and Sheng Li. 2021. 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*, pages 2767–2776, Online. Association for Computational Linguistics.

Taoyu Su, Xinghua Zhang, Jiawei Sheng, Zhenyu Zhang, and Tingwen Liu. 2024. Loginmea: Local-to-global interaction network for multi-modal entity alignment. In *ECAI 2024*, pages 1173–1180. IOS Press.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proc. VLDB Endow.*, 13(12):2326–2340.

Yuwei Wan, Ying Liu, Zheyuan Chen, Chong Chen, Xinyu Li, Fu Hu, and Michael Packianather. 2024. Making knowledge graphs work for smart manufacturing: Research topics, applications and prospects. *Journal of Manufacturing Systems*, 76:103–132.

Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang, and Shuqiang Jiang. 2020. Logo-2k+: A large-scale logo dataset for scalable logo classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6194–6201.

Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. 2017. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE international conference on computer vision*, pages 1202–1211.

Yuejia Xiang, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Zhenxi Lin, and Yefeng Zheng. 2021. OntoEA: Ontology-guided entity alignment via joint knowledge graph embedding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1117–1128, Online. Association for Computational Linguistics.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3857–3866, New York, NY, USA. Association for Computing Machinery.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2024a. 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*, SIGIR '24, page 91–101, New York, NY, USA. Association for Computing Machinery.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2024b. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. *Preprint*, arXiv:2404.09468.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 2024c. 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, Torino, Italia. ELRA and ICCL.

A Image Embedding Methods in MMKGs

Semantically Ambiguous Images. Current MMKG models typically embed images as vectors and combine them with embeddings from other modalities (e.g., text) to create richer entity representations. These embedding methods generally follow two approaches:

- Global feature extraction: Methods like Convolutional Neural Networks (CNNs) generate fixed-size global feature vectors for entire images (Li et al., 2022). While efficient for large-scale datasets, they often fail to capture fine-grained details.

- Local feature extraction: Approaches such as Vision Transformers (ViTs) divide images into patches and embed each patch individually, enabling finer-grained feature extraction and

improved alignment with text (Dosovitskiy et al., 2020). However, these methods are computationally intensive and heavily reliant on image quality and alignment effectiveness.

Challenges in Handling Specific Image Types. However, both approaches face challenges when processing certain types of image, where standard embedding methods fail to capture essential semantic features.

- Sparse-Semantic Images (e.g., brand logos): These images contain limited visual information, often featuring simple geometric shapes or elements. While they may carry critical domain knowledge, existing models struggle to extract distinctive embeddings, reducing their effectiveness (Su et al., 2024; Wang et al., 2020).

- Rich-Semantic Images (e.g., abstract artwork): These images are visually and semantically complex, including intricate scenes, interactions, or artistic expressions. Current embedding methods often struggle to fully capture these semantic relationships, leading to significant information loss (Wilber et al., 2017).

B Standardizing and Aligning Original Images

To generate textual descriptions based on images, we require the original images from the datasets. However, many existing works only provide image embeddings (vectors) without the raw images. For those that do provide raw images, naming conventions for image files vary significantly. Some use Wikidata URLs, others use DBpedia URLs, and some rely on YAGO entity names. This inconsistency, especially with YAGO entity names that often include special characters such as `\`, `/`, or `:`, creates challenges in aligning image with entity names. Many operating systems are unable to handle filenames containing such characters, further complicating the alignment process and subsequent experiments.

To address this, we first standardize the naming conventions for raw images in the datasets. Concretely, we align all entities using their Wikidata IDs (QIDs). The QIDs consist only of alphanumeric characters, which are compatible with all operating systems, facilitating future reproduction and extensions. Additionally, QIDs serve as a bridge between entities and their Wikipedia pages, enabling us to download supplementary images and metadata (e.g., timestamps) from Wikipedia for dataset enrichment. Details are summarized in Table 1.

MKG-W. We found that original images for different entities were stored in folders named after the entities , but many special characters (e.g., `\`, `/`, or `:`) were missing. Additionally, the image filenames within these folders lacked any recognizable pattern. To address this, we used the dataset's provided mapping file, which links DBpedia URLs to Wikidata URLs, to identify the corresponding entity names and QIDs for each entity.

Next, we removed all special characters from both the extracted entity names and the folder names containing the original images to facilitate matching. Once the matching was complete, we had the following information for each sample: entity name, QID, and original images. Finally, we renamed all images using the format `qid_idx` and consolidated them into a single folder for use in subsequent experiments.

MKG-Y. We followed a similar process as in MKG-W (see above). The original images were stored in folders named after the entities, but the filenames lacked a consistent naming convention. Unlike MKG-W, the original dataset did not provide a mapping file between DBpedia and Wikidata URLs. However, it did include a mapping between DBpedia URLs and sample names.

Using the DBpedia URLs, we accessed the corresponding DBpedia pages and leveraged `sameAs` links to locate the corresponding Wikidata pages and obtain the QIDs. We then matched the folders containing raw images to their respective entities and renamed the images using the `qid_idx` format. Finally, all renamed images were consolidated into a single folder for subsequent use.

DB15K. The original paper (Liu et al., 2019) did not provide downloadable images, only image embeddings and URLs for the images. As a result, we re-downloaded the images using the provided links. Each sample had 100 links from Google Images, approximately 35 from Bing Images, and 50 from Yahoo Image Search. According to the original paper, the top 20 images from each search engine (for a total of 60 images per entity) should be downloaded.

However, some links were no longer valid. To ensure fairness in reproducing the results, we sequentially downloaded up to 20 images from each search engine. If fewer than 20 valid images were

available, we continued downloading from subsequent links until 20 images were obtained per search engine, maintaining the original dataset's image count of 60 per sample.

After downloading the images, we used the DBpedia URLs to access the DBpedia pages, followed `sameAs` links to locate the corresponding Wikidata pages, and obtained the QIDs for each sample. Finally, we renamed the images using the `qid_idx` format and consolidated them into a single folder for subsequent experiments.

C Our Datasets Structure

During processing, images in each batch were encoded as tensors using a processor and then passed through the model to generate textual descriptions. The generated descriptions and their corresponding filenames were saved to an output file, providing semantically meaningful textual data for subsequent MMKG tasks. A summary of the final output files is provided in Table 3.

Key	Description
`id`	Unique identifier for each image
`page_url`	URL of the Wikipedia page
`image_url`	URL of the image file
`table_data`	Metadata of the image
`- Description`	Brief description of the image
`- Date`	Date associated with the image
`- Author`	Author or creator of the image
`- Formatted_Date`	Standardized date format
`image_blip2_detail`	Detailed textual description

Table 3: Each image in the dataset includes unique identifiers, source URLs, metadata (e.g., date, author), and BLIP-2-generated textual descriptions.

The dataset provides detailed information for each image, such as URLs, metadata, and automatically generated textual descriptions. Each entry consists of a unique identifier (`id`), links to the corresponding Wikipedia page (`page_url`) and image file (`image_url`), metadata extracted from the image (`table_data`), and textual descriptions generated by BLIP-2 (`image_blip2_detail`). The metadata includes various attributes such as date (`Formatted_Date`), author, and resolution, which are crucial for MMKG research.

D Implementation Details

Our experiments use the default hyperparameters for each Baseline model to ensure fair comparisons. All experiments were conducted on a Linux server equipped with a single NVIDIA H100 GPU. To generate textual descriptions from images, we used the BLIP-2 model, with English as the output language. The maximum generated text length is limited to 100 words. On average, each image generated 20 words, ranging from 15 to 25 words. The generated text is then embedded into vectors using `BERT-base-uncased`.

E Main Result

The main results are shown in Table 4, which summarizes the link prediction performance of four models (MMRNS, MyGO, NativE, and AdaMF) across three datasets (MKG-W, MKG-Y, and DB15K) under different settings.

Model performance is evaluated using rank-based metrics, including Mean Reciprocal Rank (MRR) and Hits@K (K = 1, 3, 10). MRR calculates the average of the reciprocal ranks of the correct answers in the predicted ranking list, while Hits@K measures the proportion of correct answers appearing within the top K predictions. Both metrics are commonly used in evaluating link prediction tasks, with higher scores indicating better model performance.

The first row for each model presents the experimental results on the original datasets, as reproduced from the original papers. "***G***" stands for ***Generate***, referring to our framework that generates textual descriptions from images. "***o***" indicates that the textual descriptions were generated from the original images provided in the dataset, while "***n***" means that the descriptions were generated from images automatically downloaded using our ***Beyond Images*** framework. ***Improvement*** represents the percentage increase (Boost = $\frac{\text{Our Result} - \text{Baseline Result}}{\text{Baseline Result}}$) in performance of the enriched datasets compared to the original datasets.

F Case Analysis: Boosting Performance with Textual Descriptions

F.1 Example 1

All images are shown in Figure 5. Triple: ***(Hot Sauce Committee Part Two, performer, Beastie Boys)***. Images (a) and (b) correspond to the head entity ***Hot Sauce Committee Part Two***, while images (c) - (h) represent the tail entity ***Beastie Boys***.

Triple: (Hot Sauce Committee Part Two, performer, Beastie Boys)

QID: (Q1933719, P175, Q214039)

Head entity's rank: correct head entity's rank improved from 13,680 to 1,330.

Model	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	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47	32.68	23.01	37.86	51.01
MMRNS (D+I+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
Improvement (↑%)	1.99%	3.70%	2.35%	2.57%	1.84%	4.11%	2.87%	2.11%	2.74%	4.49%	3.35%	3.34%
MMRNS (D+I+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
Improvement (↑%)	3.13%	4.68%	2.91%	3.26%	2.79%	4.69%	3.21%	2.44%	2.11%	3.36%	3.06%	2.72%
MMRNS (D+I+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**
Improvement (↑%)	**3.50%**	**5.20%**	**3.21%**	**3.62%**	**3.07%**	**5.21%**	**3.56%**	**2.71%**	**3.04%**	**5.01%**	**3.71%**	**3.69%**
MyGO	36.10	29.78	38.54	47.75	38.51	33.39	39.03	47.87	37.72	30.08	41.26	52.21
MyGO (D+I+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
Improvement (↑%)	3.01%	3.61%	2.88%	2.10%	2.91%	4.01%	2.17%	2.14%	2.97%	4.81%	2.69%	2.92%
MyGO (D+I+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
Improvement (↑%)	3.28%	4.97%	3.12%	2.99%	3.42%	5.03%	3.01%	2.81%	2.78%	3.81%	2.53%	1.97%
MyGO (D+I+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**
Improvement (↑%)	**3.66%**	**5.51%**	**3.48%**	**3.35%**	**3.80%**	**5.60%**	**3.31%**	**3.13%**	**3.31%**	**5.36%**	**2.99%**	**3.27%**
NativE	36.58	29.56	39.65	48.94	39.04	34.79	40.89	46.18	37.16	28.01	41.36	54.13
NativE (D+I+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
Improvement (↑%)	2.16%	3.38%	1.98%	2.02%	1.52%	3.33%	2.55%	1.83%	4.10%	2.92%	2.72%	1.81%
NativE (D+I+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
Improvement (↑%)	2.72%	3.80%	3.02%	2.72%	1.81%	3.83%	2.99%	2.71%	3.08%	2.72%	2.41%	1.66%
NativE (D+I+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**
Improvement (↑%)	**3.02%**	**4.21%**	**3.32%**	**3.01%**	**2.01%**	**4.25%**	**3.32%**	**2.98%**	**4.52%**	**3.25%**	**3.02%**	**2.02%**
AdaMF	35.85	29.04	39.01	48.42	38.57	34.34	40.59	45.76	35.14	25.30	41.11	52.92
AdaMF (D+I+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
Improvement (↑%)	2.98%	3.84%	1.96%	1.90%	3.16%	2.99%	2.12%	1.43%	3.02%	3.71%	2.87%	2.71%
AdaMF (D+I+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
Improvement (↑%)	3.77%	4.51%	1.94%	2.70%	3.84%	4.44%	3.20%	2.23%	2.01%	3.08%	2.49%	2.49%
AdaMF (D+I+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**
Improvement (↑%)	**4.21%**	**5.02%**	**2.15%**	**3.02%**	**4.25%**	**4.95%**	**3.52%**	**2.46%**	**3.35%**	**4.12%**	**3.20%**	**3.00%**

Table 4: Link prediction results of four models across three datasets. "***D***" represents entity descriptions, "***I***" denotes image embeddings, "***G(o)***" refers to textual descriptions generated from original images, and "***G(n)***" corresponds to textual descriptions from newly downloaded images. "***H@n***" stands for "Hits at n." The "***Improvement (↑%)***" indicates the performance gain of the best-performing model (highlighted in bold) over the Baseline model.

Tail entity's rank: correct tail entity's rank improved from 11,435 to 4,628.

F.2 Example 2

All images are shown in Figure 6. Triple: ***(Her Harem, cast member, Carroll Baker)***. Images (a) - (c) correspond to the head entity ***Her Harem***, while images (d) - (m) represent the tail entity ***Carroll Baker***.

Triple: (Her Harem, cast member, Carroll Baker)

QID: (Q3819142, P161, Q233891)

Head entity's rank: correct head entity's rank improved from 10,177 to 8,611.

Tail entity's rank: correct tail entity's rank improved from 571 to 72.

F.3 Example 3

All images are shown in Figure 7. Triple: ***(World (The Price of Love), performer, New Order)***. Images (a) correspond to the head entity ***World (The Price of Love)***, while images (b) - (f) represent the tail entity ***New Order***.

Triple: (World (The Price of Love), performer, New Order)

QID: (Q8035321, P175, Q214990)

Head entity's rank: correct head entity's rank improved from 12,528 to 2,622.

Tail entity's rank: correct tail entity's rank improved from 10,185 to 2,591.

(a) Q1933719_1: “The cover of beastboys hot sauce committee part two”.

(b) Q1933719_2: “The scene shows a group of men walking on a bridge”.

(c) Q214039_1: “three men are leaning on a stair railing”.

(d) Q214039_2: “The logo for beastie boys is shown in black and white”.

(e) Q214039_3: “two men are standing on stage with a microphone”.

(f) Q214039_4: “two men in black jackets are on stage singing”.

(g) Q214039_5: “a man in a red suit and hat is singing on stage”.

(h) Q214039_6: “a man in a suit and tie singing”.

Figure 5: Triple: ***(Hot Sauce Committee Part Two, performer, Beastie Boys)***. Images (a) and (b) correspond to the head entity ***Hot Sauce Committee Part Two***, while images (c) - (h) represent the tail entity ***Beastie Boys***.

(a) Q3819142_1: "the poster for the movie harem".

(b) Q3819142_2: "the italian flag is shown on a clapperboard".

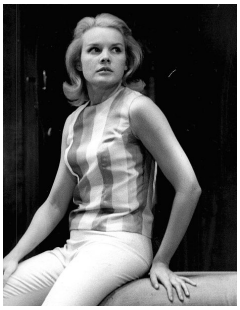
(c) Q3819142_3: "two masks and a clapper board on a black background".

(d) Q233891_1: "a black and white photo of a woman with long blonde hair".

(e) Q233891_2: "a black background with a white tv screen".

(f) Q233891_3: "marilyn monroe in a black and white photo".

(g) Q233891_4: "a man and woman in western attire sit on a horse".



(h) Q233891_5: "a woman in a striped top sits on a bench".

(i) Q233891_6: "a woman in a fur coat sits on a white fur rug".

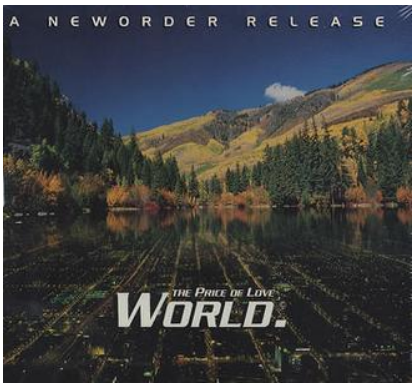
(j) Q233891_7: "a woman is standing in a shower".

(k) Q233891_8: "the scene shows a man and woman talking to each other".

(l) Q233891_9: "a woman in a white dress is standing on a stage in front of a large ship".

(m) Q233891_10: "a star on the hollywood walk of fame for carroll baker".

Figure 6: Triple: ***(Her Harem, cast member, Carroll Baker)***. Images (a) - (c) correspond to the head entity ***Her Harem***, while images (d) - (m) represent the tail entity ***Carroll Baker***.



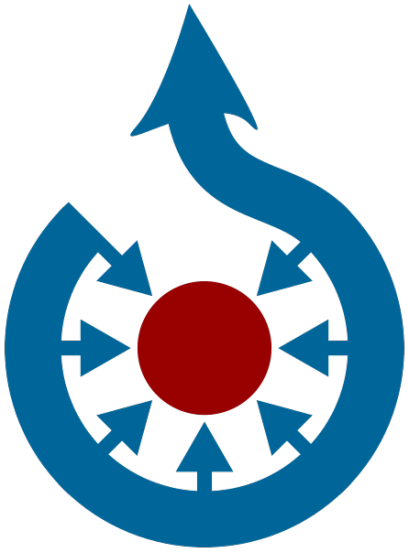
(a) Q8035321_1: "the cover of the world album".

(b) Q214990_1: "four black and white photos of four men".

(c) Q214990_2: "a group of men are on stage with guitars and drums".

(d) Q214990_3: "a band is performing on stage with a large screen behind them".

(e) Q214990_4: "a blue and white wave symbol".



(f) Q214990_5: "a blue and red logo with arrows pointing in different directions".

Figure 7: Triple: ***(World (The Price of Love), performer, New Order)***. Images (a) correspond to the head entity ***World (The Price of Love)***, while images (b) - (f) represent the tail entity ***New Order***.