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

Supplementary Material

1 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., symbolic 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 paintings): 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).

2 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

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.

	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

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.

3 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 2.

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

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

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 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`.

5 Computational Cost

Our method works directly with existing images in MMKG datasets and achieves 3-4% improvement (Figure 3 in the paper), validating our main hypothesis(RQ1). Retrieving new images from the internet is optional and was only used for the ablation study(RQ2). While external images can bring additional gains (Table 3 and Figure 3 in the paper), they are not required for the method to be effective.

Image retrieval and captioning is a one-time preprocessing step, similar to standard MMKG setups. As detailed in the supplementary (Sec. 2&4), generating descriptions for 100K images takes 1 hour (MKG-W/Y) or 8 hours (DB15K) on a single H100 GPU. The output text files are <200MB and reusable.

Once generated, this enriched data can be reused across models and tasks without extra cost. As shown in the table, training time increases by only 7-30 minutes, while performance improves, demonstrating a favorable cost-benefit trade-off.

6 Image-to-Text Generation Models

"***Blip2-flan-t5-xxl***" 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

Model	Group	Variant	Time
MMRNS	I+T	Baseline	143.82
	T+G	G(o)	146.27
		G(n)	145.71
		G(o+n)	146.02
		Fusion	146.32
	I+T+G	G(o)	175.25
		G(n)	175.31
		G(o+n)	175.29
		Fusion	175.76
MyGO	I+T	Baseline	77.56
	T+G	G(o)	80.23
		G(n)	80.16
		G(o+n)	79.99
		Fusion	80.02
	I+T+G	G(o)	102.04
		G(n)	102.03
		G(o+n)	102.04
		Fusion	102.21
NativE	I+T	Baseline	46.94
	T+G	G(o)	47.92
		G(n)	47.94
		G(o+n)	47.91
		Fusion	47.85
	I+T+G	G(o)	56.05
		G(n)	56.04
		G(o+n)	56.10
		Fusion	56.15
AdaMF	I+T	Baseline	33.38
	T+G	G(o)	35.24
		G(n)	35.26
		G(o+n)	35.29
		Fusion	35.48
	I+T+G	G(o)	42.57
		G(n)	42.50
		G(o+n)	42.53
		Fusion	42.63

Table 3: Training time (minutes) for each model across groups and variants.

overhead and eliminating the need for extensive retraining.

"***Git-large-coco***" model adopts a generative Transformer architecture that directly converts CLIP image tokens into textual descriptions. Fine-tuned on the COCO image captioning dataset, it is well-suited for generating coherent and relevant captions. Its end-to-end training allows it to extract visual details and express them in natural language effectively.

"***Llava-v1.5-7b***" model integrates a pre-trained CLIP vision encoder with a large language model connected through a lightweight projection layer. Fine-tuned on visual instruction data, it generates detailed, context-aware descriptions conditioned on both the image and the user prompt. It leverages the generative and reasoning capabilities of the language model to produce high-quality image-grounded text.

7 Human-in-the-Loop Consistency Check Tool Design

7.1 Tool Design

7.1.1 Goal and Operating Mode

The tool runs locally in a browser and supports offline use. It presents, for each QID, the LLM summary and the associated image set, enables fast annotation, and persists every change at the record level. An optional "LLM Assist" switch illustrates a hypothetical workflow but does not call external services.

7.1.2 Data and Indexing

Input is a single JSON file (not JSONL) whose records contain `entity_name`, `entity_qid`, `dbpedia_url`, `wikidata_url`, and `images_descriptions`. Images live in a local directory with filenames `{QID}_{index}.jpg`. The application builds in-memory indices `QID` → record and `QID` → image paths.

7.1.3 Setup Wizard

On first run the tool opens a setup page. The user selects the text JSON path and the image folder, then loads the dataset. The tool validates paths, creates or refreshes thumbnail cache at `.cache/thumbnails/`, and, if absent, clones the original JSON into `dataset_curated.json`. Global options (paths, theme, and key bindings) are saved to `settings.json`.

7.1.4 Interface and Interaction

A top bar shows the title and status badges (Reviewed, Pending, Mismatch, Uncertain) that also act as filters, plus a "Jump to QID" box. The main view has two panels: the left panel shows entity metadata and a collapsible, scrollable text description; the right panel shows an image grid with hover metadata and lightbox zoom. Each image supports "Remove from this entity" for soft deletion, including batch selection. A verdict area offers Match, Mismatch, and Uncertain; Mismatch reasons (wrong entity, irrelevant, low quality, other); notes; and `Save` or `Save & Next`. A Settings panel controls theme, key mappings, caches, and a guarded option for physical deletion. The LLM Assist toggle exposes a non-functional demo flow.

7.1.5 Persistence Semantics

Autosave is enabled by default (timer and on-record switch), and manual Save is available. On save, all changes for the current QID are written to `dataset_curated.json`. The file mirrors the original schema and adds a `_curation` object:

```
"_curation": {
  "verdict": "match|mismatch|uncertain",
 "reason": "wrong_entity|irrelevant|low_quality|other
  "notes": "string",
 "removed_images": ["Q21197_3.jpg", "..."],
  "annotator": "optional_id",
  "last_updated": "ISO-8601",
  "modified": true
}
```

Soft deletion only records filenames and hides them in the UI; it does not touch disk files. Physical deletion is an opt-in setting with confirmation.

7.1.6 Export and Collaboration

Two export modes are available: *Export All* (full curated JSON) and *Export Modified* (only entries with `modified=true`). Exports never copy images. The tool supports single-user mode, local network viewing with single-writer saving, and sharded annotation by QID with later merge based on timestamp or annotator priority.

7.1.7 Performance and Usability

The tool uses lazy loading and lookahead prefetching (current and the next and previous three records) for thumbnails. Filtering and jump rely on in-memory indices to avoid full re-rendering. Default shortcuts are `A` for Match, `D` for Mismatch, `U` for Uncertain, `S` for Save, `Ctrl+Enter` for Save & Next, arrows for navigation, and `G` for jump. Light and dark themes are available, and long text can be folded.

7.1.8 Local Deployment

A minimal stack uses `streamlit`, `pillow`, and optionally `pandas`. Run with `streamlit run app.py`. The tool opens in the default browser and performs all I/O on local disk.

7.2 Verification Protocol

7.2.1 Annotation Unit and Presentation

One unit equals one QID. The view contains the LLM summary and the complete image set for that entity. Annotators judge the alignment based on the main semantics of the images, such as subject, place or object, and salient symbols.

7.2.2 Labels and Rubric

Match: the text captures the main visual semantics of the image set; minor wording differences are allowed.
Mismatch: the text conflicts with or is unrelated to the images.
Uncertain: evidence is insufficient to decide due to low resolution, ambiguity, or incomplete context.
An optional Partial category can be mapped to Mismatch with the reason set to “other:partial” if finer granularity is needed.

7.2.3 Operations and Error Handling

For fixable issues, edit the text field `images.images_t5_descriptions` and save. For irrelevant or low-quality images, apply soft deletion so they no longer appear for that entity. For important or disputed cases, add notes and schedule a second pass; the final decision is recorded after review.

7.2.4 Persistence, Recovery, and Autosave

On Save or autosave, the verdict, reason, notes, removed images, and any text edits are written under `_curation` in `dataset_curated.json`. On the next launch, the tool loads the curated file if present; otherwise it creates a new curated copy from the original JSON. Unsaved changes in the browser session are discarded on close; the tool warns on exit to prevent accidental loss.

7.2.5 Export and Incremental Consumption

Exports contain only JSON. *Export All* writes the full curated file with `_curation`. *Export Modified* writes only entries with `modified=true`, which supports incremental downstream processing. Images remain in the original directory; thumbnail cache files serve only the interface.

7.2.6 Edge Cases and Quality Control

If text or images are missing, the tool highlights the issue and still allows a verdict with notes. Path changes can be fixed in the settings and the index refreshed. Files that do not follow the naming pattern `{QID}_*.ext` are ignored and listed in logs. If physical deletion is enabled, the tool deletes files listed under `removed_images` after a confirmation step. For sample-based audits, we recommend stratified sampling of about ten percent, double annotation with Cohen’s κ reporting, and adjudication of conflicts. Comparing pre-fusion and post-fusion consistency rates quantifies the benefit of LLM fusion.

8 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.

MMRNS[1] (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.

MyGO[2] (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[3] (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[4] (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.

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***” indi-

[1] https://github.com/quqxui/MMRNS
[2] https://github.com/zjukg/MyGO
[3] https://github.com/zjukg/NATIVE
[4] https://github.com/zjukg/AdaMF-MAT

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.

	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 + 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 + 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 + 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%
MMRNS + LLM 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**
Improvement (↑%)	**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	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 + 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 + 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 + 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%
MyGO + LLM 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**
Improvement (↑%)	**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	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 + 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 + 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 + 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%
NativE + LLM 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**
Improvement (↑%)	**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	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 + 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 + 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 + 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%
AdaMF + LLM 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**
Improvement (↑%)	**6.11%**	**7.76%**	**3.54%**	**4.21%**	**5.10%**	**6.82%**	**4.37%**	**3.77%**	**4.56%**	**6.63%**	**4.33%**	**3.63%**

cates 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.

9 Case Analysis: Boosting Performance with Textual Descriptions

9.1 Example 1

All images are shown in Figure 1. 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.

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

9.2 Example 2

All images are shown in Figure 2. 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.

9.3 Example 3

All images are shown in Figure 3. 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.

References

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.

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.

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.

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.

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.

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) 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 1: 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".

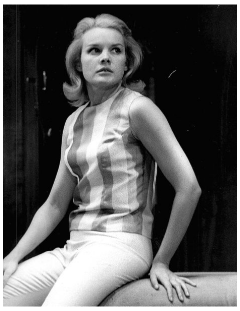

(c) Q3819142_3: "two masks and a clapper board on a black background".


(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".

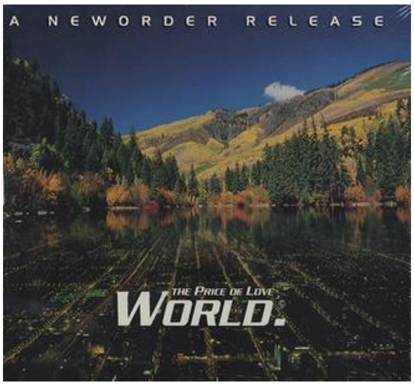
(j) Q233891_7: "a woman is standing in a shower".


(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 2: 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***.



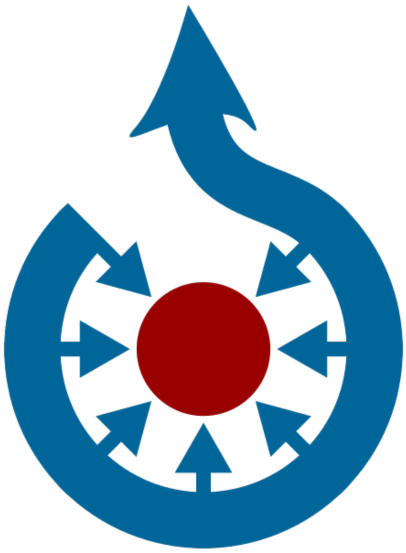
(a) Q8035321_1: "the cover of the world album".

(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 3: 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***.