Multi-modal Knowledge Graphs: Evolution, Methods, and Opportunities

Anonymous ACL submission

Abstract

Knowledge Graphs (KGs) are pivotal in advancing AI applications, and their extension into multi-modal dimensions (MMKGs) is opening new avenues for innovation. This survey systematically defines MMKGs, charts their construction progress, and analyzes existing MMKG-related tasks. We provide detailed task definitions, evaluation benchmarks, and insights into significant breakthroughs, while also discussing current challenges and highlighting emerging trends in the field.

1 Introduction

Knowledge Graphs (KGs) play a critical role in structuring long-tail knowledge and serve as foundational elements in many successful AI systems (Hogan et al., 2022). While traditional KGs offer considerable benefits, their focus on single-modality knowledge restricts their applicability to multi-modal tasks. For example, scenarios with complex visual details are difficult to enhance solely through text-based knowledge, highlighting the need for Multi-Modal Knowledge Graphs (MMKGs) that incorporate symbols from other modalities (e.g., Vision). This integration offers a viable strategy for overcoming the limitations of traditional KGs and broadening their capabilities, as illustrated in Fig. 1. Within this paper, we first trace the progression from conventional KGs to MMKGs, noting the evolving focus within the semantic web community. We then carefully explore the impact of multi-modal techniques on KGs, discussing both their current state and future prospects. Detailed analysis covers methodological developments within each task and benchmarks key areas, enabling effective comparison across tasks. Focusing primarily on research from the past three years, we also includes a discussion on the recent advancements in Large Language Models (LLMs), exploring their synergies with the aforementioned topics. In summary, this survey aims to offer a comprehensive, insightful overview of the MMKG field, offering deep insights into the evolving landscape and guiding future studies.

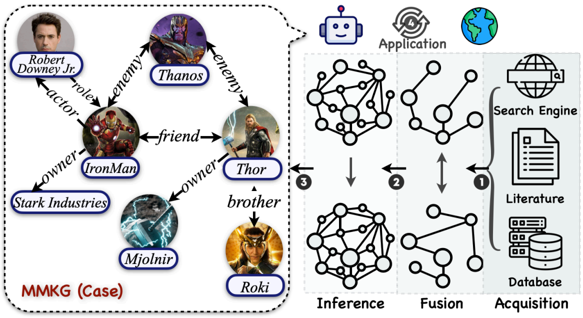


Figure 1: Roadmap for Multi-Modal Knowledge Graph construction and application.

2 Preliminaries

Knowledge Graphs. KGs represent entities and their relationships in a graph structure, where nodes symbolize real-world entities or atomic values (attributes), and edges denote relations. Knowledge in KG is often captured in triples, with an ontology-based schema defining basic entity classes and their relations in a taxonomic structure. A KG is defined as $\mathcal{G} = \mathcal{E}, \mathcal{R}, \mathcal{T}$, with entities $\mathcal{E}$, relations $\mathcal{R}$, and statements $\mathcal{T}$. Statements include relational fact triples (h, r, t) (i.e., $\mathcal{T}_\mathcal{R} = \mathcal{E} \times \mathcal{R} \times \mathcal{E}$), where h is the head entity, r is the relation, and t is the tail entity, or attribute triples (e, a, v) (i.e., $\mathcal{T}_\mathcal{A} = \mathcal{E} \times \mathcal{A} \times \mathcal{V}$), where e is an entity, a is an attribute, and v is the attribute's value. v can be literals such as strings or dates and may include metadata like labels and textual definitions.

Ontology. Within the semantic web community, ontologies serve as KG schemas with key features including: ***(i)*** Hierarchical classes, often termed as concepts; ***(ii)*** Properties that specify the terms used in relations; ***(iii)*** Hierarchies involving both concepts and relations; ***(iv)*** Constraints, including the domain and range of relations, as well as class disjointness; ***(v)*** Logical expressions that encompass relation composition.

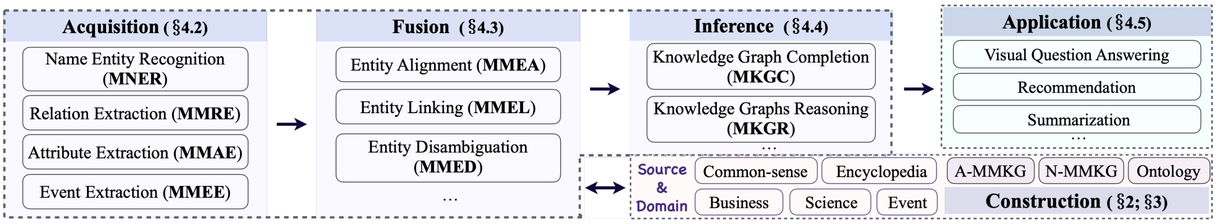


Figure 2: Comprehensive Overview of Multi-modal Knowledge graph research. Due to space constraints and task overlaps, we focus on the most representative sub-tasks in each category (Acquisition, Fusion & Inference, Application) to maximize relevant content coverage. Additional content is analyzed in the Appendix[3].

Languages like RDF, RDFS[1], and OWL[2] introduce built-in vocabularies to capture these knowledge elements, ensuring richer semantics and superior quality (Horrocks, 2008) with predicates like *rdfs:subClassOf* denoting concept subsumption.

Multi-modal Knowledge Graphs. A KG qualifies as multi-modal (MMKG) when it contains knowledge symbols expressed in multiple modalities, which can include, but are not limited to, text, images, sound, or video. This survey distinguishes between two MMKG representation methods, A-MMKG and N-MMKG, as inspired by Zhu et al. (2022a), where A-MMKGs treat images as entity attributes, and N-MMKGs allow images to stand as independent entities with direct relationships:

- **A-MMKG** utilizes multi-modal data (e.g., images) as specific **attribute values** for entities or concepts, with $\mathcal{T}_{\mathcal{A}} = \mathcal{E} \times \mathcal{A} \times (\mathcal{V}_{KG} \cup \mathcal{V}_{MM})$, where $\mathcal{V}_{KG}$ and $\mathcal{V}_{MM}$ are values of KG and multi-modal data, respectively.
- **N-MMKG** treats multi-modal data as KG **entities**, with $\mathcal{T}_{\mathcal{R}} = (\mathcal{E}_{KG} \cup \mathcal{E}_{MM}) \times \mathcal{R} \times (\mathcal{E}_{KG} \cup \mathcal{E}_{MM})$, separating typical KG entities ($\mathcal{E}_{KG}$) from multi-modal entities ($\mathcal{E}_{MM}$).

Given the convenience in data access and similarity to traditional KGs, A-MMKG forms the basis for most current applications or methods in MMKG research, as elaborated in § 4.3 and § 4.4.

MMKG Construction. We outline two principal paradigms following Zhu et al. (2022b):

(i) Annotating Images with Symbols from a KG, which prioritizes the extraction of visual entities/concepts, relations, and events, crucial for the dynamic creation of KGs like scene and event graphs (Ma et al., 2022). This approach, however, faces challenges in representing infrequent (i.e., long-tail) multi-modal knowledge, primarily due to the recurrent depiction of common real-world entities across diverse contexts. The use of supervised methods further compounds these challenges, as they are inherently constrained by the finite scope of pre-existing labels. Moreover, those systems demands substantial pre-processing, including the formulation of specific rules, the creation of predetermined entity lists, and the application of pretrained detectors and classifiers, all of which pose significant scalability challenges (Li et al., 2020a).

Typical construction methods for most of the current MMKG is *(ii)* Grounding KG Symbols to Images, which involves: entity grounding (i.e., associating entities with corresponding images from online sources (Oñoro-Rubio et al., 2019)), concept grounding (i.e., selecting diverse, representative images for visual concepts and abstracting common visual features), and relation grounding (i.e., choosing images that semantically mirror the relation of the input triples). Nevertheless, considering the scale factor, this paradigm currently poses the principal challenge in large-scale MMKG construction.

3 MMKG Evolution

In Appendix A.2.2 and Tab. 1, we provide a detailed exposition of MMKG-related work prior to **2021**, initially centered on defining MMKG concepts and frameworks. Recently, the focus in the MMKG community has shifted from **Construction** to **Refinement** and **Application**. Specifically, Peng et al. (2022) explore image quality control in MMKG construction through an Image Refining Framework that uses clustering for de-duplication and noise reduction, taps into Wikidata for entity descriptions, and relies on a pre-trained model to gauge image-text similarity, discarding images below a certain relevance threshold. In MMKG construction, accurately aligning concepts with their corresponding images is crucial. The challenge arises from distinguishing between visualizable

[1] RDF Schema, https://www.w3.org/TR/rdf-schema/

[2] Web Ontology Language, https://www.w3.org/TR/owl2-overview/

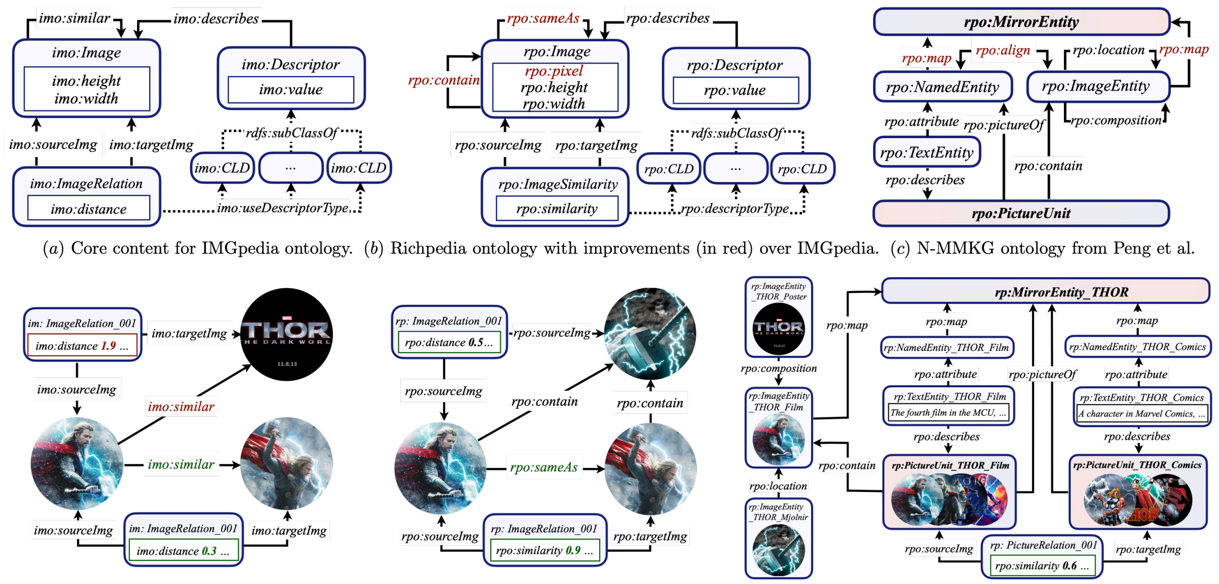
[3] For a focused discussion, most **method references**, **detailed descriptions** and **benchmarks** are organized in the Appendix for readers interested in tracing the original sources.

Table 1: Overview of various MMKGs, detailing their publish (Pub.) time, types, scale, data sources, and supported (Sup.) tasks, where symbol $*$ indicates the inclusion of triple-level multi-modal information within the MMKG. Not that only part of the Sup. tasks are listed that have been experimentally validated in original studies, although MMKGs have a wider potential task range. The key distinctions among nodes, entities, and concepts are based on their representation: entities typically correspond directly to real-world object names, nodes include both these entities and textual elements (Alberts et al., 2020) like Wikipedia articles, and concept is a further decomposition of entity where each entity has multiple concepts, corresponding to different aspects such as "*culture*", "*geography*", and "*history*" (Zhang et al., 2023a; Zha et al., 2023). Besides, this table primarily lists MMKGs in general visual multi-modal scenarios, excluding other event-based or domain-specific MMKGs like ManipMob-MMKG (Song et al., 2023c), which focuses on indoor scenes. Abbreviations used: **Data source**: CN (ConceptNet); DBP (DBpedia); Freebase (FB); VG (VisualGenome); WP (Wikipedia); WN (WordNet); WD (Wikidata); Wikimedia (WM); Web Search Engine (WSE); YG (YAGO). **Tasks**: Image Classification (IMGC); Cross-Modal Retrieval (CMR); Object Detection (OD); Scene Graph Generation (SGG); Visual Question Anwering (VQA); Concept Understanding (CU); Multi-modal Knowledge Graph Completion (MKGC), Knowledge Graph Reasoning (MKGR), Entity Alignment (MMEA), Entity Linking (MMEL) and Information Extraction (MMIE).

Pub. Time	MMKGs	Types	Scale (#nodes / #images)	Data Sources	Sup. Tasks
2013-12	NEIL (Chen et al., 2013)	N-MMKG	1152 (classes) / 300K	WN / Image WSE	OD, etc.
2014-09	ImageNet (Russakovsky et al., 2015)	A-MMKG	21K (classes) / 3.2M	WN / Image WSE	IMGC, OD, etc.
2016-02	VisualGenome (Krishna et al., 2017)	A-MMKG	35 (classes) / 108K	WN / MS COCO / YFCC (Thomee et al., 2016)	SGG, VQA, etc.
2016-09	WN9-IMG (Xie et al., 2017)	A-MMKG	6.5K (entities) / 14K	WN / ImageNet	MKGC
2017-01	ImageGraph (Liu et al., 2017)	A-MMKG	15K (entities) / 837K	FB / Image WSE	CMR
2017-10	IMGpedia (Ferrada et al., 2017)	N-MMKG	2.6M (entities) / 15M	DBP / WM Commons	CMR
2019-03	MMKG (Liu et al., 2019b)	A-MMKG	45K (entities) / 37K	FB / DBP / YG / Image WSE	MMEA, MKGC
2020-07	GAIA (Li et al., 2020a)	N-MMKG	457K (entities) / NA	FB / GeoNames / News Websites	MMIE
2020-08	VisualSem (Alberts et al., 2020)	N-MMKG	90K (nodes) / 938K	WP / WN / ImageNet	CMR
2020-09	DBP-DWY-Vis (Liu et al., 2021)	A-MMKG	178K (entities) / 117K	WP / DBP15k (Sun et al., 2017) / DWY15K (Guo et al., 2019)	MMEA
2020-12	Richpedia (Wang et al., 2020)	N-MMKG	2.8M (entities) / 2.9M	WD / WM / Image WSE	MMKG Querying
2021-06	RESIN (Wen et al., 2021)	N-MMKG	51K (events) / NA	WD / News Websites	MMIE
2022-10	MKG-W&Y (Xu et al., 2022b)	A-MMKG	30K (entities) / 29K	OpenEA (Sun et al., 2020c) / Image WSE	MKGC
2022-10	MarKG (Zhang et al., 2023b)	A-MMKG	11K (entities) / 76K	WD / Image WSE	MKGR
2023-02	Multi-OpenEA (Li et al., 2023l)	A-MMKG	920K (entities) / 2.7M	OpenEA / Image WSE	MMEA
2023-03	UKnow (Gong et al., 2023)	N-MMKG	1.4M (entities) / 1.1M	WP / Image WSE	MKGC, CMR
2023-07	UMVM (Chen et al., 2023f)	A-MMKG	238K (entities) / 205K	DBP-DWY-Vis / Multi-OpenEA	MMEA
2023-08	AspectMMKG (Zhang et al., 2023a)	A-MMKG	2.3K (entities) / 645K	WP / Image WSE	MMEL
2023-10	TIVA-KG (Wang et al., 2023h)	A-MMKG$*$	440K (entities) / 1.7M	CN / Image WSE	MKGC
2023-11	MMpedia (Wu et al., 2023b)	A-MMKG	2.7M (entities) / 19.5M	DBP / Image WSE	MKGC
2023-12	VTKGs (Lee et al., 2023)	A-MMKG$*$	43K (entities) / 460K	CN / WN / UnRel (Peyre et al., 2017) / VRD (Lu et al., 2016) HICO-DET (Chao et al., 2018) / VisKE (Sadeghi et al., 2015)	MKGC
2023-12	M^{2}ConceptBase (Zha et al., 2023)	A-MMKG	152K (concepts) / 951K	Wukong (Gu et al., 2022) / Baidu Encyclopedia	VQA, CU

concepts (VCs), like "*dog*", which have clear visual representations, and non-visualizable concepts (NVCs), such as "*mind*" or "*texture*", which lack direct visual counterparts. Jiang et al. (2022) introduce a visual concept classifier that identifies VCs and NVCs, utilizing ImageNet instances to exemplify the former. This initial binary classification is just a preliminary step, as the main challenge in MMKG construction involves selecting representative images for entities, potentially through clustering methods like K-means or spectral clustering (Zhu et al., 2022b). Building upon this, Zhang et al. (2023a) introduce **AspectMMKG**, enriching MMKGs by associating entities with aspect-specific images and refining image selection with a trained model. Besides, Wu et al. (2023b) present **MMpedia**, a scalable, high-quality MMKG constructed via a pipeline that leverages DBpedia (Auer et al., 2007) to filter NVCs and refine entity-related images using textual and type information.

Toward addressing complex multi-modal scenarios and further automating MMKG construction, Gong et al. (2023) introduce **UKnow**, a unified knowledge protocol that categorizes N-MMKG triples into five unit types: in-image, in-text, cross-image, cross-text, and image-text. They establish a pipeline convert existing datasets into UKnow's format, simplifying the creation of new datasets from existing image-text pairs. Additionally, Zha et al. (2023) present **M^{2}ConceptBase**, a multi-modal conceptual MMKG framework. Initially, they extract candidate concepts from textual descriptions in image-text pairs and refine them using rule-based filters. These concepts are then aligned with corresponding images and detailed descriptions through context-aware multi-modal symbol grounding. For concepts not fully grounded, GPT-3.5-Turbo generates supplementary descriptions. Wang et al. (2023h) investigate the impact of different modalities in Link Prediction via **TIVA-KG**, an MMKG covering text, image, video, and audio. Built upon the foundation of ConceptNet (Speer et al., 2017), TIVA-KG supports **triplet grounding** (i.e., associating a common-sense triplet with tangible representations like images). Similarly, Lee et al. (2023) propose **VTKGs**, where images are



(a) Core content for IMGpedia ontology. (b) Richpedia ontology with improvements (in red) over IMGpedia. (c) N-MMKG ontology from Peng et al.

(d) MMKG example using the IMGpedia ontology. (e) Example using the Richpedia ontology. (f) MMKG example using the ontology similiar to Peng et al.

Figure 3: Representative N-MMKG ontologies and corresponding MMKG examples using those ontologies.

attached to both entities/triplets, and each entity/relation is accompanied by textual descriptions.

N-MMKG Ontology: URI prefixes are crucial in ontologies, uniquely identifying classes and properties and ensuring compliance with RDF standards. Standard prefixes (e.g., *rdf*, *rdfs*, *owl*) ensure cross-domain consistency, while custom ones (e.g., *imo* for IMGpedia and *rpo* for Richpedia) bring in domain-specific nuances. Fig. 3 visualizes the evolutionary trajectory of MMKG ontologies (detailed in Appendix A.2.2), highlighting the unique challenges N-MMKGs face: ***(i)*** An individual entity may exhibit multiple visual representations (i.e., varied aspects). ***(ii)*** Efficiently extracting information from visual modalities across entities is crucial. ***(iii)*** Development of diverse multi-modal representation methods can extend from entity-level to relation and triple-level, as explored in recent works (Wang et al., 2023h; Lee et al., 2023).

4 Multi-modal Knowledge Graph Tasks

4.1 MMKG Representation Learning

Late Fusion methods emphasize modality interactions and feature aggregation just prior to output generation (Fig. 9 (a)). MKGRL-MS (Wang et al., 2022b) crafts unique single-modal embeddings, employing multi-head self-attention to determine each modality's contribution to semantic composition and **sum** the weighted multi-modal features for MMKG entity representation. MMKRL (Lu et al., 2022b) learns cross-modal embeddings in a unified translational semantic space, merging them through **concatenation**. DuMF (Li et al., 2022c) applies a bilinear layer for feature projection and an attention block for modality preference learning in each track, integrating features via a **gate network**.

Early Fusion methods integrate multi-modal feature at an initial stage, enabling full modality interactions for complex reasoning (Fig. 9 (b)). Fang et al. (2023b) first normalizes entity modalities into a unified embedding using an MLP, then refines them by contrasting with perturbed negative samples. MMRotatH (Wei et al., 2023b) utilizes a gated encoder to merge textual and structural data, filtering irrelevant information within a rotational dynamics-based KGE framework. Recent studies (Lee et al., 2023) utilize (V)PLMs like BERT and ViT for multi-modal data integration. They format graph structures, text, and images into sequences or dense embeddings compatible with LMs, thereby utilizing the LMs' reasoning capabilities and the knowledge embedded in their parameters to support tasks such as Multi-modal Link Prediction.

4.2 MMKG Acquisition

As the first step in MMKG construction (Fig. 1), MMKG Acquisition (or Extraction), involves integrating multi-modal data from sources like search engines or public databases to enhance existing KGs or develop new MMKGs.

Named Entity Recognition (NER) identifies and classifies named entities in text into categories like persons, organizations, and locations. For example, in the sentence "*Apple Inc. is founded by Steve Jobs in California*", NER models would identify "*Apple Inc.*" as an organization, "*Steve Jobs*" as a person, and "*California*" as a location. Multi-modal Named Entity Recognition (MNER) extends this

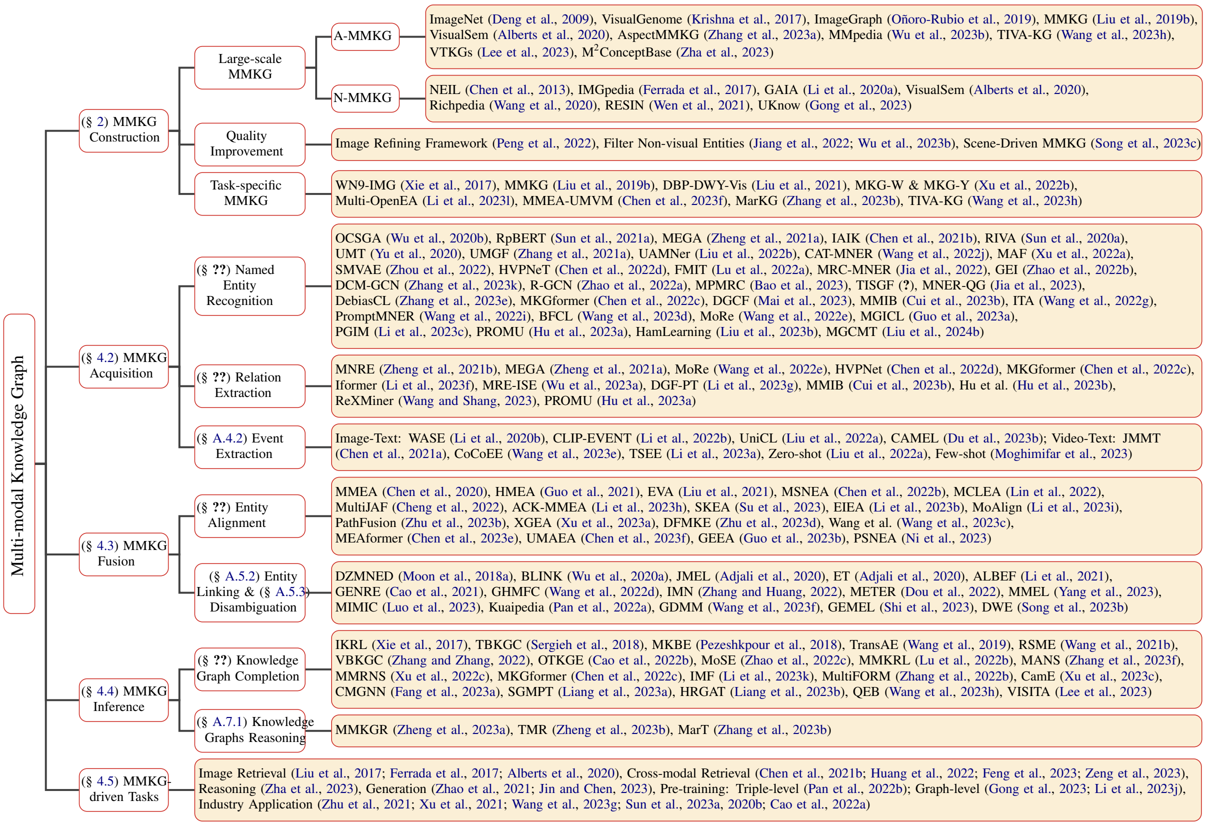


Figure 4: Taxonomy of the Multi-modal Knowledge Graph Realm, with the "Multi-modal" prefix omitted for clarity.

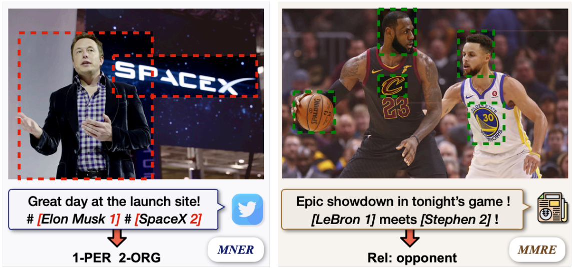


Figure 5: Illustrative examples demonstrating the application scenarios for MNER (left) and MMRE (right).

process by incorporating visual information(Chen et al., 2023d). Similarly, Relation Extraction (RE) detects and classifies semantic relationships between entities within text identifying a "*founded by*" relationship between "*Apple Inc.*" and "*Steve Jobs*" in the same sentence. Multi-modal Relation Extraction (MMRE) uses visual cues to enrich these analyses, proving effective in contexts like news articles where text accompanies images or videos. For further details, see Appendix A.4.1 and Fig. 5.

MNER. ***(i)* BiLSTM-based Methods** (Moon et al., 2018b) primarily employ a modality attention network to combine text and image features, incorporating a visual attention gate within LSTM to better recognize named entities in social media posts. ***(ii)* PLM-based Methods** (Yu et al., 2020) modifies the standard PLM (e.g., BERT) structure for MNER by adding a Transformer layer for enhanced text representation and a cross-modal Transformer for visual integration. Some of them find visual inputs effective in identifying entity types but not spans, leading to the inclusion of a text-only module for more accurate entity span detection. ***(iii)* Special Cases:** Certain studies address unique challenges in MNER. For example, DebiasCL (Zhang et al., 2023e) focuses on bias mitigation in MNER through a visual object density-guided hard sample mining strategy and a debiased contrastive loss.

MMRE. Zheng et al. (2021b) first demonstrate how multi-modal data can bridge semantic gaps and improve social media text analysis. Building on this, works like (Zheng et al., 2021a; Wu et al., 2023a) introduce a textual-visual relation alignment method that synchronizes sentence parsing trees with visual scene graphs for more precise MMRE. Similarly, PLM-based methods (Chen et al., 2022d; Li et al., 2023g) employ approaches akin to those in MNER.

4.3 MMKG Fusion

The proliferation of heterogeneous data across the Internet has led to the creation of numerous independent MMKGs. Integrating these from diverse sources is essential, making MMKG fusion a critical stage in large-scale MMKG construction. Entity Alignment (EA) is pivotal for KG integration, aiming to match identical entities across different KGs by leveraging their relational, attributive, and literal (surface) features. Multi-Modal Entity Alignment (MMEA) extend EA by incorporating visual data from MMKGs, linking each entity with images to improve accuracy (Liu et al., 2019b).

MMEA: Current MMEA research falls into two streams based on underlying motivation. ***(i)* Exploring better cross-KG modality feature fusion:** Techniques include extending MMKG representation from Euclidean to hyperbolic space for better geometric interpretation (Guo et al., 2021); assigning different importance to each modality via a global-level attention (Liu et al., 2021) or instance-level transformer (Chen et al., 2023e; Li et al., 2023i; Wang et al., 2024a) mechanism; strengthening this process through contrastive learning (Lin et al., 2022); leveraging visual cues to guide relational feature learning and prioritize valuable attributes for alignment (Chen et al., 2022b).

***(ii)* Analyzing practical limitations and challenges in MMKG alignment:** The inherent incompleteness of visual data in MMKGs is a challenge as many entities lack images. Additionally, the intrinsic ambiguity of visual images impacts alignment quality due to multiple visual aspects per entity, as detailed in § 2. Wang et al. (2023c) address image-type mismatches in aligned multi-modal entities by using pre-defined ontologies and an image type classifier to filter out incongruent images. Chen et al. (2023f) explore the effects of training noise from high rates of missing modalities. Guo et al. (2023b) tackle the issue of dangling entities, which lack counterparts in the target KG, by generating new entities conditionally or unconditionally using a mutual variational autoencoder.

4.4 MMKG Inference

MMKG data inherently contain missing elements, errors, and contradictions, making inference a critical task for MMKG completion (Fig. 1). The goal of MKGC is to enrich the relational triple set $\mathcal{T}_R$ in A-MMKGs by identifying missing relational triples among entities and relations, potentially using attribute triples $\mathcal{T}_A$. Specifically, Entity Prediction identifies missing head or tail entities in queries $(h, r, ?)$ or $(?, r, t)$; Relation Prediction pinpoints missing relations in $(h, ?, t)$; Triple Classification determines the truth of triples (h, r, t). Notably, most current MKGC efforts focus on Entity Prediction, also known as Link Prediction.

MKGC: Mainstream MKGC approaches primarily follow two paths: ***(i)* Embedding-based Approaches** evolve from traditional KGE techniques (Bordes et al., 2013), adapting them to include multi-modal data, thus forming multi-modal entity embeddings. These approaches include: **Modality Fusion** methods (Wilcke et al., 2023), integrating multi-modal embeddings of entities with their structural embeddings for triple plausibility estimation, utilizing methods like multiple TransE-based scoring functions (Xie et al., 2017), transformer framework (Lee et al., 2023) or optimal transport (Cao et al., 2022b) for modal interaction. **Modality Ensemble**, where separate models for different modalities combine outputs for final predictions (Zhao et al., 2022c; Li et al., 2023k). **Modality-aware Negative Sampling**, generating false triples to improve model discernment between accurate and erroneous KG triples (Zhang and Zhang, 2022; Xu et al., 2022c). ***(ii)* Fine-Tuning based Approaches** leverage pre-trained Transformer models like BERT and VisualBERT (Li et al., 2019) to utilize their deep multi-modal understanding for MKGC. These methods transform MMKG triples into token sequences for PLMs (Liang et al., 2022), treating MKGC tasks as classification problems where PLMs encode KG information and predict masked entities (Chen et al., 2022c).

4.5 MMKG-driven Tasks

In this section, we explore several key directions where MMKGs have shown notable impact in downstream task applications.

Retrieval. As discussed in § 2, several MMKGs could naturally support retrieval related tasks like ImageGraph (Liu et al., 2017), IMGpedia (Ferrada et al., 2017), and VisualSem (Alberts et al., 2020).

MMKG-driven Cross-modal Retrieval methods like MKVSE (Feng et al., 2023), which scores intra- and inter-modal relations in MMKGs using WordNet path similarity and co-occurrence correlations (Fig. 6), improving Image-Text Retrieval through GNN-based multi-modal embeddings.

Reasoning & Generation. Multi-modal reasoning and generation tasks often demand specialized knowledge, including long-tail information that ex-

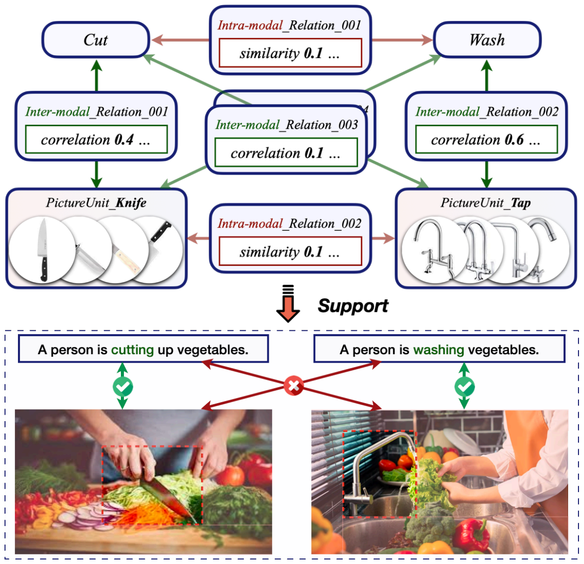


Figure 6: We illustrates the MMKG-supported Image-Text Retrieval process (Feng et al., 2023). For simplicity, all URI prefixes and certain relations (*sourceImg* and *targetImg*) from the *PictureRelation* (*Inter-modal_Relation* and *Intra-modal_Relation*) entity are omitted. This entity's values indicate intra-modal path similarities or inter-modal co-occurrence correlations, essential for training a model (e.g., multi-modal GCN) to produce knowledgeable image or text representations. Note: In cases of multiple images within a picture unit, mean pooling is used for a unified feature representation.

ceeds common experiences. KGs are crucial in these scenarios, serving as structured repositories for such diverse knowledge. However, there exists a notable gap between KGs and multi-modal tasks, as current methods frequently depend on indirect approaches like modal transformation for knowledge representation, retrieval, and interaction in multi-modal contexts. This becomes problematic in tasks requiring visual common sense, leading to multi-modal hallucinations (Fig. 7). Recent works (Zha et al., 2023) demonstrate that MMKGs can effectively bridge this gap. Specifically, Zha et al. (2023) introduce M^2ConceptBase (detailed in § 2), a conceptual MMKG that improves VQA performance by retrieving multi-modal concept descriptions and crafting instructions to refine answers with MLLMs.

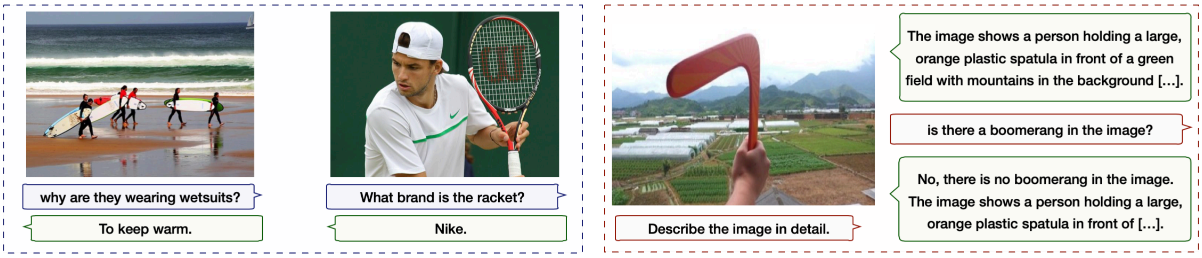
MMKG Pre-training. Current VLMs, which primarily pre-train on basic image-text pairs, often overlook extensive intermodal knowledge connections, prompting the development of MMKG-based methods: ***(i)*** **Triple-level** methods treat triples as independent knowledge units, implicitly embedding the (h, r, t) structure into the VLM's embedding space. For example, Pan et al. (2022b) integrate knowledge-based objectives into the CLIP framework using MMKGs like Visual Genome (Krishna et al., 2017) and VisualSem. They use the CLIP encoder to process textual and visual entities and their relationships, fusing them via a multi-modal Transformer. This approach enhances CLIP's training with a triple-based loss function, improving performance across various multi-modal tasks. ***(ii)*** **Graph-level** methods capitalize on the structural connections among entities in a global MMKG. By selectively gathering multi-modal neighbor nodes around each entity featured in the training corpus, they apply techniques such as GNNs or concatenation to incorporate knowledge during the pre-training process (Gong et al., 2023; Li et al., 2023j).

Industry Application. E-commerce greatly benefits from Multi-modal Product KGs (MMPKGs) as depicted in Fig. 8. This integration supports key applications such as product management, comparison, and recommendation. The K3M (Zhu et al., 2021) framework utilizes MMPKGs to improve product representations through techniques like masked object prediction, masked language reconstruction, and link prediction, enriching pre-training and integration of multi-modal knowledge.

5 Future Directions

MMKGs, along with KGs, aim to address the lack of long-tail knowledge in various tasks, reflecting real-world patterns and human experiences. Current research optimistically assumes that an infinitely expansive MMKG could nearly encapsulate all relevant world knowledge, potentially solving multi-modal challenges effectively. However, important questions remain: How can we acquire **ideal multi-modal knowledge**? What should the ideal MMKG feature, and can it **mirror the human brain's sophisticated understanding of the world**? Additionally, does MMKG provide unique benefits over the **knowledge capabilities of LLMs**? Addressing these questions is crucial as we continue to delve into this field.

MMKG Construction & Acquisition. ***(i)*** Aligning locally extracted triples from multiple images with large-scale KGs (Lee et al., 2023) not only extends the coverage of image quantity but also incorporates the extensive knowledge scale characteristic. ***(ii)*** Refining and aligning fine-grained knowledge within MMKGs is crucial. An ideal MMKG should be hierarchical, containing deep,



(a) LLMs (e.g., BLIP-2) applied in multi-modal reasoning tasks when lacking visual background knowledge.

(b) LLMs (e.g., MiniGPT-4) applied in multi-modal generative tasks when lacking fine-grained visual knowledge alignment.

Figure 7: Examples of limited cross-modal knowledge alignment ability in current MLLMs (Zha et al., 2023), specifically (a) BLIP-2 (Li et al., 2023e) and (b) MiniGPT-4 (Zhu et al., 2023c), leading to hallucinations.

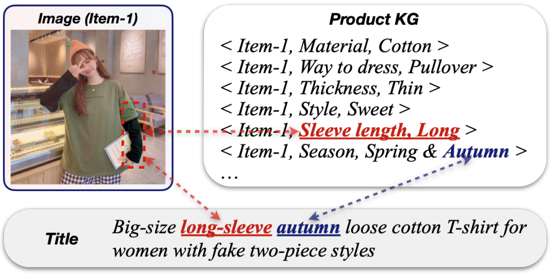


Figure 8: Illustration of multi-modal product data in MMPKGs (Zhu et al., 2021), representing each product with a title, an image, and relevant parts of the Product Knowledge Graph (PKG) through triples such as *(item, property, value)*. MMPKG pre-training enhances VLMs by improving visual grounding and domain-specific multi-modal knowledge comprehension in E-commerce.

detailed layers of abstract multi-modal knowledge, allowing a single image to represent multiple concepts. Moreover, segmentation represents an advanced requirement for grounding to reduce background noise in visual modalities, pushing towards **segmentation-level and multi-grained** MMKGs as a key future direction. ***(iii)*** Efficiency in MMKG storage and utilization: Despite traditional KGs' efficiency in storing vast knowledge with minimal parameters, MMKGs require more storage space, presenting challenges in data storage and task application. ***(iv)*** Quality control: Regular updates are crucial given the dynamic nature of knowledge, with future directions focusing on efficiently resolving multi-modal knowledge conflicts and updates.

MMKG for Tasks. Several factors limit the use of MMKGs across diverse tasks: ***(i)*** Non-Uniform Organization and Ontology: Current MMKGs lack a standardized format and vary in focus and knowledge domains, primarily catering to encyclopedic or trivia knowledge (Gong et al., 2023), with commonsense and scientific MMKGs (Lee et al., 2023) being notably rare. Moreover, the abstract knowledge often cannot be visualized, limiting practical use (Wu et al., 2023b). ***(ii)*** Data Timeliness and Completeness: Inadequacies in these areas heighten the risk of multi-modal hallucinations. ***(iii)*** Comparative Advantages of LLMs and MLLMs: Noted for their generalizability and AGI potential (Zhang et al., 2024), LLMs and MLLMs complement MMKGs' interpretability and flexibility. The development, maintenance, and operational costs of MMKGs, coupled with industry feedback, shape perceptions of their practicality. ***(iv)*** Rich Semantic MMKG Construction: MMKGs extend beyond traditional formats by transforming multi-modal datasets into semantically enriched structures through task-specific pipelines, utilizing existing KGs as bases. This method enhances MLLM training with structured inputs and contributes semantically rich datasets to the MMKG community. ***(v)*** Reconstruction of Multi-Modal Tasks with LLM: By leveraging the text understanding and generation capabilities of LLMs, multimodal tasks can be restructured. Converting KG-driven multi-modal tasks into in-MMKG tasks (e.g., MKGC and MMEA) can improve domain integration (Pahuja et al., 2024). ***(vi)*** MMKG MoE: The Mixed of Expert (MoE) architecture enhances LLM applications by selectively routing inputs through GateNet for efficient expert selection (Ismail et al., 2023). Proposing a specialized MMKG library for domains like biology could mirror this approach, optimizing MMKG utilization and integration with dynamic allocation efficiency.

6 Conclusion

This paper presents a thorough review of MMKG construction evolution, analyzing key tasks and methods relevant to the field. By providing detailed benchmarking, we aim to create a systematic blueprint of the domain, establishing it as a valuable resource for researchers either currently engaged in or planning to enter this area. Ultimately, this review serves as a foundational guide, mapping the trajectory and potential of MMKG research and highlighting future opportunities.

7 Limitations

In this study, we provide an overview of problems, methods, and opportunities for multi-modal knowledge graph research. We discuss related surveys in Appendix A.1 and will continue adding more related approaches with more detailed analysis. Despite our best efforts, there may be still some limitations that remain in this paper.

References & Methods. Due to the page limit, we may have omitted some important references and cannot afford all the technical details. Our Literature Collection Methodology is shared in Appendix A.1. We primarily review cutting-edge methods from the past three years (mostly in 2023), sourced from major conferences and journals like ACL, EMNLP, NAACL, CVPR, NeurIPS, ICLR, and arXiv, etc., and we will continue to update our review with the latest research.

Benchmarks. Most of the benchmarks mentioned (e.g., Tab. 5 and Tab. 7) are gathered and categorized from the experimental part of mainstream works. In order to help readers quickly understand the tasks' goals and formats from a unified perspective, the definition and boundary of each task may not be accurate enough.

Empirical Conclusions. We provide detailed comparisons and discussions on in-MMKG methods in § 4, listing some promising future directions in § 5. All conclusions are based on empirical analysis of existing works, which may not capture a broader perspective. As the field evolves, we will update our findings to reflect the latest developments.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *ECIR (1)*, volume 12035 of *Lecture Notes in Computer Science*, pages 463–478. Springer.

Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, and Huan Liu. 2023. Can knowledge graphs reduce hallucinations in llms? : A survey. *CoRR*, abs/2311.07914.

Houda Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2020. Visualsem: a high-quality knowledge graph for vision and language. *CoRR*, abs/2008.09150.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L. Yuille, Trevor Darrell, Jitendra Malik, and Alexei A. Efros. 2023. Sequential modeling enables scalable learning for large vision models. *CoRR*, abs/2312.00785.

Xigang Bao, Mengyuan Tian, Zhiyuan Zha, and Biao Qin. 2023. MPMRC-MNER: A unified MRC framework for multimodal named entity recognition based multimodal prompt. In *CIKM*, pages 47–56. ACM.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. *CoRR*, abs/2401.05856.

Matthias Baumgartner, Luca Rossetto, and Abraham Bernstein. 2020. Towards using semantic-web technologies for multi-modal knowledge graph construction. In *ACM Multimedia*, pages 4645–4649. ACM.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*, pages 1247–1250. ACM.

Stephen Bonner, Ian P. Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L. Hamilton. 2022. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings Bioinform.*, 23(6).

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Anna Breit, Simon Ott, Asan Agibetov, and Matthias Samwald. 2020. Openbiolink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, 36(13):4097–4098.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *ICLR*. OpenReview.net.

Xianshuai Cao, Yuliang Shi, Jihu Wang, Han Yu, Xinjun Wang, and Zhongmin Yan. 2022a. Cross-modal knowledge graph contrastive learning for machine learning method recommendation. In *ACM Multimedia*, pages 3694–3702. ACM.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022b. OTKGE: multi-modal knowledge graph embeddings via optimal transport. In *NeurIPS*.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *WACV*, pages 381–389. IEEE Computer Society.

Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. 2021a. Joint multimedia event extraction from video and article. In *EMNLP (Findings)*, pages 74–88. Association for Computational Linguistics.

Chen Chen, Yufei Wang, Bing Li, and Kwok-Yan Lam. 2022a. Knowledge is flat: A seq2seq generative framework for various knowledge graph completion. In *COLING*, pages 4005–4017. International Committee on Computational Linguistics.

Chen Chen, Yufei Wang, Aixin Sun, Bing Li, and Kwok-Yan Lam. 2023a. Dipping plms sauce: Bridging structure and text for effective knowledge graph completion via conditional soft prompting. In *ACL (Findings)*, pages 11489–11503. Association for Computational Linguistics.

Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021b. Multimodal named entity recognition with image attributes and image knowledge. In *DASFAA (2)*, volume 12682 of *Lecture Notes in Computer Science*, pages 186–201. Springer.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2023b. LION : Empowering multimodal large language model with dual-level visual knowledge. *CoRR*, abs/2311.11860.

Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020. MMEA: entity alignment for multi-modal knowledge graph. In *KSEM (1)*, volume 12274 of *Lecture Notes in Computer Science*, pages 134–147. Springer.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022b. Multimodal siamese network for entity alignment. In *KDD*, pages 118–126. ACM.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023c. Unveiling the siren's song: Towards reliable fact-conflicting hallucination detection. *CoRR*, abs/2310.12086.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022c. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR*, pages 904–915. ACM.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022d. Good visual guidance make A better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *NAACL-HLT (Findings)*, pages 1607–1618. Association for Computational Linguistics.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: extracting visual knowledge from web data. In *ICCV*, pages 1409–1416. IEEE Computer Society.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2014. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, pages 2035–2042. IEEE Computer Society.

Yong Chen, Xinkai Ge, Shengli Yang, Linmei Hu, Jie Li, and Jinwen Zhang. 2023d. A survey on multimodal knowledge graphs: Construction, completion and applications. *Mathematics*, 11(8):1815.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z. Pan, Wenting Song, and Huajun Chen. 2023e. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *ACM Multimedia*, pages 3317–3327. ACM.

Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z. Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023f. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *ISWC*, volume 14265 of *Lecture Notes in Computer Science*, pages 121–139. Springer.

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z. Pan, and Huajun Chen. 2023g. DUET: cross-modal semantic grounding for contrastive zero-shot learning. In *AAAI*, pages 405–413. AAAI Press.

Zhuo Chen, Wen Zhang, Yufeng Huang, Mingyang Chen, Yuxia Geng, Hongtao Yu, Zhen Bi, Yichi Zhang, Zhen Yao, Wenting Song, Xinliang Wu, Yi Yang, Mingyi Chen, Zhaoyang Lian, Yingying Li, Lei Cheng, and Huajun Chen. 2023h. Teleknowledge pre-training for fault analysis. In *ICDE*, pages 3453–3466. IEEE.

Bo Cheng, Jia Zhu, and Meimei Guo. 2022. Multijaf: Multi-modal joint entity alignment framework for multi-modal knowledge graph. *Neurocomputing*, 500:581–591.

Jian Cheng, Kaifang Long, Shuang Zhang, Tian Zhang, Lianbo Ma, Shi Cheng, and Yinan Guo. 2023a. Text-image scene graph fusion for multi-modal named entity recognition. *IEEE Transactions on Artificial Intelligence*.

Siyuan Cheng, Xiaozhuan Liang, Zhen Bi, Huajun Chen, and Ningyu Zhang. 2023b. Multi-modal protein knowledge graph construction and applications (student abstract). In *AAAI*, pages 16190–16191. AAAI Press.

Wikimedia Commons. 2012. Wikimedia commons. *Retrieved June*, 2.

Congcong Ge and Xiaoze Liu and Lu Chen and Baihua Zheng and Yunjun Gao. 2021. Largeea: Aligning entities for large-scale knowledge graphs. *Proc. VLDB Endow.*, 15(2):237–245.

UniProt Consortium. 2019. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

Hejie Cui, Xinyu Fang, Zihan Zhang, Ran Xu, Xuan Kan, Xin Liu, Yue Yu, Manling Li, Yangqiu Song, and Carl J. Yang. 2023a. Open visual knowledge extraction via relation-oriented multimodality model prompting. *CoRR*, abs/2310.18804.

Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2023b. Enhancing multimodal entity and relation extraction with variational information bottleneck. *CoRR*, abs/2304.02328.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *CoRR*, abs/2401.06066.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2018. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *ICLR (Poster)*. OpenReview.net.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society.

Ludovic Denoyer and Patrick Gallinari. 2006. The wikipedia xml corpus. In *ACM SIGIR Forum*, volume 40, pages 64–69. ACM New York, NY, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Xin Luna Dong. 2023. Generations of knowledge graphs: The crazy ideas and the business impact. *CoRR*, abs/2308.14217.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *CoRR*, abs/2312.09979.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, pages 18145–18155. IEEE.

Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. 2023a. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *CoRR*, abs/2311.01487.

Zilin Du, Yunxin Li, Xu Guo, Yidan Sun, and Boyang Li. 2023b. Training multimedia event extraction with generated images and captions. In *ACM Multimedia*, pages 5504–5513. ACM.

Quan Fang, Xiaowei Zhang, Jun Hu, Xian Wu, and Changsheng Xu. 2023a. Contrastive multi-modal knowledge graph representation learning. *IEEE Trans. Knowl. Data Eng.*, 35(9):8983–8996.

Quan Fang, Xiaowei Zhang, Jun Hu, Xian Wu, and Changsheng Xu. 2023b. Contrastive multi-modal knowledge graph representation learning. *IEEE Trans. Knowl. Data Eng.*, 35(9):8983–8996.

Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2022. Molecular contrastive learning with chemical element knowledge graph. In *AAAI*, pages 3968–3976. AAAI Press.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. 2023c. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, pages 1–12.

Duoduo Feng, Xiangteng He, and Yuxin Peng. 2023. MKVSE: multimodal knowledge enhanced visual-semantic embedding for image-text retrieval. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(5):162:1–162:21.

Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. 2017. Imgpedia: A linked dataset with content-based analysis of wikimedia images. In *ISWC (2)*, volume 10588 of *Lecture Notes in Computer Science*, pages 84–93. Springer.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: A new dataset and A baseline. In *MM*, pages 993–1001. ACM.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. 2012. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107.

Congcong Ge, Xiaoze Liu, Lu Chen, Baihua Zheng, and Yunjun Gao. 2021. Make it easy: An effective end-to-end entity alignment framework. In *SIGIR*, pages 777–786. ACM.

Yuxia Geng, Jiaoyan Chen, Yuhang Zeng, Zhuo Chen, Wen Zhang, Jeff Z. Pan, Yuxiang Wang, and Xiaoliang Xu. 2023. Prompting disentangled embeddings for knowledge graph completion with pre-trained language model. *CoRR*, abs/2312.01837.

Biao Gong, Xiaoying Xie, Yutong Feng, Yiliang Lv, Yujun Shen, and Deli Zhao. 2023. Uknow: A unified knowledge protocol for common-sense reasoning and vision-language pre-training. *CoRR*, abs/2302.06891.

Dihong Gong and Daisy Zhe Wang. 2017. Extracting visual knowledge from the web with multimodal learning. In *IJCAI*, pages 1718–1724. ijcai.org.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *NeurIPS*.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2023. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. *CoRR*, abs/2311.13314.

Aibo Guo, Xiang Zhao, Zhen Tan, and Weidong Xiao. 2023a. MGICL: multi-grained interaction contrastive learning for multimodal named entity recognition. In *CIKM*, pages 639–648. ACM.

Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing*, 461:598–607.

Lingbing Guo, Zhuo Chen, Jiaoyan Chen, and Huajun Chen. 2023b. Revisit and outstrip entity alignment: A perspective of generative models. *CoRR*, abs/2305.14651.

Lingbing Guo, Zequn Sun, and Wei Hu. 2019. Learning to exploit long-term relational dependencies in knowledge graphs. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 2505–2514. PMLR.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2022. Knowledge graphs. *ACM Comput. Surv.*, 54(4):71:1–71:37.

Ian Horrocks. 2008. Ontologies and the semantic web. *Communications of the ACM*, 51(12):58–67.

Xuming Hu, Junzhe Chen, Aiwei Liu, Shiao Meng, Lijie Wen, and Philip S. Yu. 2023a. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In *ACM Multimedia*, pages 5185–5194. ACM.

Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S. Yu. 2023b. Multimodal relation extraction with cross-modal retrieval and synthesis. In *ACL (2)*, pages 303–311. Association for Computational Linguistics.

Ningyuan Huang, Yash R. Deshpande, Yibo Liu, Houda Alberts, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2022. Endowing language models with multimodal knowledge graph representations. *CoRR*, abs/2206.13163.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023a. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2023b. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310.

Vassilis N. Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xiangxiang Zeng, and George Karypis. 2020. Drkg - drug repurposing knowledge graph for covid-19. https://github.com/gnn4dr/DRKG/.

Aya Abdelsalam Ismail, Sercan O Arik, Jinsung Yoon, Ankur Taly, Soheil Feizi, and Tomas Pfister. 2023. Interpretable mixture of experts. *Transactions on Machine Learning Research*.

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. MNER-QG: an end-to-end MRC framework for multimodal named entity recognition with query grounding. In *AAAI*, pages 8032–8040. AAAI Press.

Meihuizi Jia, Xin Shen, Lei Shen, Jinhui Pang, Lejian Liao, Yang Song, Meng Chen, and Xiaodong He. 2022. Query prior matters: A MRC framework for multimodal named entity recognition. In *ACM Multimedia*, pages 3549–3558. ACM.

Xueyao Jiang, Ailisi Li, Jiaqing Liang, Bang Liu, Rui Xie, Wei Wu, Zhixu Li, and Yanghua Xiao. 2022. Visualizable or non-visualizable? exploring the visualizability of concepts in multi-modal knowledge graph. In *DASFAA (1)*, volume 13245 of *Lecture Notes in Computer Science*, pages 180–187. Springer.

Lingyun Jin and Jingqiang Chen. 2023. Self-supervised opinion summarization with multi-modal knowledge graph. *Journal of Intelligent Information Systems*, pages 1–18.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678. IEEE Computer Society.

Zeeshan Khan, C. V. Jawahar, and Makarand Tapaswi. 2022. Grounded video situation recognition. In *NeurIPS*.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C Rindflesch. 2012. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160.

Bosung Kim, Taesuk Hong, Youngjoong Ko, and Jungyun Seo. 2020. Multi-task learning for knowledge graph completion with pre-trained language models. In *COLING*, pages 1737–1743. International Committee on Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. *CoRR*, abs/2304.02643.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. 2023. VISTA: visual-textual knowledge graph representation learning. In *EMNLP (Findings)*, pages 7314–7328. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880. Association for Computational Linguistics.

Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Dehai Min, Yongrui Chen, and Guilin Qi. 2023a. Three stream based multi-level event contrastive learning for text-video event extraction. In *EMNLP*. Association for Computational Linguistics.

Jinxu Li, Qian Zhou, Wei Chen, and Lei Zhao. 2023b. Enhanced entity interaction modeling for multi-modal entity alignment. In *KSEM (2)*, volume 14118 of *Lecture Notes in Computer Science*, pages 214–227. Springer.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023c. Prompting chatgpt in MNER: enhanced multimodal named entity recognition with auxiliary refined knowledge. In *EMNLP (Findings)*, pages 2787–2802. Association for Computational Linguistics.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023d. Prompting chatgpt in MNER: enhanced multimodal named entity recognition with auxiliary refined knowledge.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023e. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022a. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pages 9694–9705.

Lei Li, Xiang Chen, Shuofei Qiao, Feiyu Xiong, Huajun Chen, and Ningyu Zhang. 2023f. On analyzing the role of image for visual-enhanced relation extraction (student abstract). In *AAAI*, pages 16254–16255. AAAI Press.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022b. Clip-event: Connecting text and images with event structures. In *CVPR*, pages 16399–16408. IEEE.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare R. Voss, Daniel Napierski, and Marjorie Freedman. 2020a. GAIA: A fine-grained multimedia knowledge extraction system. In *ACL (demo)*, pages 77–86. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020b. Cross-media structured common space for multimedia event extraction. In *ACL*, pages 2557–2568. Association for Computational Linguistics.

Qian Li, Shu Guo, Cheng Ji, Xutan Peng, Shiyao Cui, Jianxin Li, and Lihong Wang. 2023g. Dual-gated fusion with prefix-tuning for multi-modal relation extraction. In *ACL (Findings)*, pages 8982–8994. Association for Computational Linguistics.

Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. 2023h. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In *WWW*, pages 2499–2508. ACM.

Qian Li, Cheng Ji, Shu Guo, Zhaoji Liang, Lihong Wang, and Jianxin Li. 2023i. Multi-modal knowledge graph transformer framework for multi-modal entity alignment. *CoRR*, abs/2310.06365.

Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. 2017. Situation recognition with graph neural networks. In *ICCV*, pages 4183–4192. IEEE Computer Society.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. 2023j. Graphadapter: Tuning vision-language models with dual knowledge graph. *CoRR*, abs/2309.13625.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023k. IMF: interactive multimodal fusion model for link prediction. In *WWW*, pages 2572–2580. ACM.

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

Yangning Li, Jiaoyan Chen, Yinghui Li, Yuejia Xiang, Xi Chen, and Hai-Tao Zheng. 2023l. Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment. In *ICASSP*, pages 1–5. IEEE.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *CoRR*, abs/2212.05767.

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

Shuang Liang, Anjie Zhu, Jiasheng Zhang, and Jie Shao. 2023b. Hyper-node relational graph attention network for multi-modal knowledge graph completion. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(2):62:1–62:21.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947.

Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. 2020. KGNN: knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, pages 2739–2745. ijcai.org.

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

Bing Liu, Tiancheng Lan, Wen Hua, and Guido Zuccon. 2023a. Dependency-aware self-training for entity alignment. In *WSDM*, pages 796–804. ACM.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *AAAI*, pages 4257–4266. AAAI Press.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *EMNLP (1)*, pages 1641–1651. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019a. Neural cross-lingual event detection with minimal parallel resources. In *EMNLP/IJCNLP (1)*, pages 738–748. Association for Computational Linguistics.

Jian Liu, Yufeng Chen, and Jinan Xu. 2022a. Multimedia event extraction from news with a unified contrastive learning framework. In *ACM Multimedia*, pages 1945–1953. ACM.

Luping Liu, Meiling Wang, Mozhi Zhang, Linbo Qing, and Xiaohai He. 2022b. Uamner: uncertainty-aware multimodal named entity recognition in social media posts. *Appl. Intell.*, 52(4):4109–4125.

Peipei Liu, Hong Li, Yimo Ren, Jie Liu, Shuaizong Si, Hongsong Zhu, and Limin Sun. 2023b. A novel framework for multimodal named entity recognition with multi-level alignments. *CoRR*, abs/2305.08372.

Peipei Liu, Gaosheng Wang, Hong Li, Jie Liu, Yimo Ren, Hongsong Zhu, and Limin Sun. 2024b. Multi-granularity cross-modal representation learning for named entity recognition on social media. *Inf. Process. Manag.*, 61(1):103546.

Weide Liu, Xiaoyang Zhong, Jingwen Hou, Shaohua Li, Haozhe Huang, and Yuming Fang. 2023c. Integrating large pre-trained models into multimodal named entity recognition with evidential fusion. *CoRR*, abs/2306.16991.

Xiaoze Liu, Junyang Wu, Tianyi Li, Lu Chen, and Yunjun Gao. 2023d. Unsupervised entity alignment for temporal knowledge graphs. In *WWW*, pages 2528–2538. ACM.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019b. MMKG: multi-modal knowledge graphs. In *ESWC*, volume 11503 of *Lecture Notes in Computer Science*, pages 459–474. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ziqiong Liu, Shengjin Wang, Liang Zheng, and Qi Tian. 2017. Robust imagegraph: Rank-level feature fusion for image search. *IEEE Trans. Image Process.*, 26(7):3128–3141.

Prisca Lo Surdo, Marta Iannuccelli, Silvia Contino, Luisa Castagnoli, Luana Licata, Gianni Cesareni, and Livia Perfetto. 2023. Signor 3.0, the signaling network open resource 3.0: 2022 update. *Nucleic Acids Research*, 51(D1):D631–D637.

Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon-Camarasa. 2023. Multiway-adapater: Adapting large-scale multi-modal models for scalable image-text retrieval. *CoRR*, abs/2309.01516.

Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, pages 852–869. Springer.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *ACL (1)*, pages 1990–1999. Association for Computational Linguistics.

Junyu Lu, Dixiang Zhang, Jiaxing Zhang, and Pingjian Zhang. 2022a. Flat multi-modal interaction transformer for named entity recognition. In *COLING*, pages 2055–2064. International Committee on Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment.

Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. 2022b. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.*, 52(7):7480–7497.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022c. Unified structure generation for universal information extraction. In *ACL (1)*, pages 5755–5772. Association for Computational Linguistics.

Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. Multi-grained multimodal interaction network for entity linking. In *KDD*, pages 1583–1594. ACM.

Shengxuan Luo and Sheng Yu. 2022. An accurate unsupervised method for joint entity alignment and dangling entity detection. In *ACL (Findings)*, pages 2330–2339. Association for Computational Linguistics.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2023. Are we falling in a middle-intelligence trap? an analysis and mitigation of the reversal curse. *CoRR*, abs/2311.07468.

Xin Lv, Yankai Lin, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Do pre-trained models benefit knowledge graph completion? A reliable evaluation and a reasonable approach. In *ACL (Findings)*, pages 3570–3581. Association for Computational Linguistics.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. 2022. MMEKG: multi-modal event knowledge graph towards universal representation across modalities. In *ACL (demo)*, pages 231–239. Association for Computational Linguistics.

Finlay MacLean. 2021. Knowledge graphs and their applications in drug discovery. *Expert opinion on drug discovery*, 16(9):1057–1069.

Weixing Mai, Zhengxuan Zhang, Kuntao Li, Yun Xue, and Fenghuan Li. 2023. Dynamic graph construction framework for multimodal named entity recognition in social media. *IEEE Transactions on Computational Social Systems*.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *CoRR*, abs/2310.02168.

George A Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

Farhad Moghimifar, Fatemeh Shiri, Reza Haffari, Yuan-Fang Li, and Van Nguyen. 2023. Few-shot domain-adaptative visually-fused event detection from text. In *FUSION*, pages 1–8. IEEE.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018a. Multimodal named entity disambiguation for noisy social media posts. In *ACL (1)*, pages 2000–2008. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018b. Multimodal named entity recognition for short social media posts. In *NAACL-HLT*, pages 852–860. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *HLT-NAACL*, pages 300–309. The Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *ACL (2)*, pages 365–371. The Association for Computer Linguistics.

Wenxin Ni, Qianqian Xu, Yangbangyan Jiang, Zongsheng Cao, Xiaochun Cao, and Qingming Huang. 2023. PSNEA: pseudo-siamese network for entity alignment between multi-modal knowledge graphs. In *ACM Multimedia*, pages 3489–3497. ACM.

Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto Gonzalez-Sanchez, and Roberto Javier López-Sastre. 2019. Answering visual-relational queries in web-extracted knowledge graphs. In *AKBC*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *CoRR*, abs/2312.05934.

Vardaan Pahuja, Weidi Luo, Yu Gu, Cheng-Hao Tu, Hong-You Chen, Tanya Y. Berger-Wolf, Charles V. Stewart, Song Gao, Wei-Lun Chao, and Yu Su. 2024. Bringing back the context: Camera trap species identification as link prediction on multimodal knowledge graphs.

Haojie Pan, Yuzhou Zhang, Zepeng Zhai, Ruiji Fu, Ming Liu, Yangqiu Song, Zhongyuan Wang, and Bing Qin. 2022a. Kuaipedia: a large-scale multi-modal short-video encyclopedia. *CoRR*, abs/2211.00732.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023a. Large language models and knowledge graphs: Opportunities and challenges. *CoRR*, abs/2308.06374.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023b. Unifying large language models and knowledge graphs: A roadmap. *CoRR*, abs/2306.08302.

Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. 2022b. Contrastive language-image pre-training with knowledge graphs. In *NeurIPS*.

Jinyoung Park, Ameen Patel, Omar Zia Khan, Hyunwoo J. Kim, and Joo-Kyung Kim. 2023. Graph-guided reasoning for multi-hop question answering in large language models. *CoRR*, abs/2311.09762.

Huang Peng, Hao Xu, Jiuyang Tang, Jibing Wu, and Hongbin Huang. 2022. Effectively filtering images for better multi-modal knowledge graph. In *APWeb/WAIM Workshops*, volume 1784 of *Communications in Computer and Information Science*, pages 10–22. Springer.

Jinghui Peng, Xinyu Hu, Wenbo Huang, and Jian Yang. 2023. What is a multi-modal knowledge graph: A survey. *Big Data Res.*, 32:100380.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL.

Bethany Percha and Russ B Altman. 2018. A global network of biomedical relationships derived from text. *Bioinformatics*, 34(15):2614–2624.

Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. 2017. Weakly-supervised learning of visual relations. In *ICCV*, pages 5189–5198. IEEE Computer Society.

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding multimodal relational data for knowledge base completion. In *EMNLP*, pages 3208–3218. Association for Computational Linguistics.

Sarah M. Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *ECCV (4)*, volume 12349 of *Lecture Notes in Computer Science*, pages 314–332. Springer.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *CoNLL Shared Task (2)*, pages 160–170. Association for Computational Linguistics.

Zhiyuan Qi, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Yuejia Xiang, Ningyu Zhang, and Yefeng Zheng. 2021. Unsupervised knowledge graph alignment by probabilistic reasoning and semantic embedding. In *IJCAI*, pages 2019–2025.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. AUTOACT: automatic agent learning from scratch via self-planning.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Vipula Rawte, Amit P. Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *CoRR*, abs/2309.05922.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252.

Fereshteh Sadeghi, Santosh Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, pages 1456–1464. IEEE Computer Society.

Tara Safavi, Doug Downey, and Tom Hope. 2022. Cascader: Cross-modal cascading for knowledge graph link prediction. *CoRR*, abs/2205.08012.

Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-sequence knowledge graph completion and question answering. In *ACL (1)*, pages 2814–2828. Association for Computational Linguistics.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV (8)*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In **SEM@NAACL-HLT*, pages 225–234. Association for Computational Linguistics.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. 2021. Entity linking meets deep learning: Techniques and solutions. *TKDE*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE*, 27(2):443–460.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2023. Generative multimodal entity linking. *CoRR*, abs/2306.12725.

Fangzhou Song, Bin Zhu, Yanbin Hao, Shuo Wang, and Xiangnan He. 2023a. CAR: consolidation, augmentation and regulation for recipe retrieval. *CoRR*, abs/2312.04763.

Shezheng Song, Shan Zhao, Chengyu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. 2023b. A dual-way enhanced framework from text matching point of view for multimodal entity linking. *CoRR*, abs/2312.11816.

Yaoxian Song, Penglei Sun, Haoyu Liu, Zhixu Li, Wei Song, Yanghua Xiao, and Xiaofang Zhou. 2023c. Scene-driven multimodal knowledge graph construction for embodied ai.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Fenglong Su, Chengjin Xu, Han Yang, Zhongwu Chen, and Ning Jing. 2023. Neural entity alignment with cross-modal supervision. *Inf. Process. Manag.*, 60(2):103174.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW*, pages 697–706. ACM.

Chengjie Sun, Weiwei Chen, Lei Lin, and Lili Shan. 2023a. Enhancing recommender system with multimodal knowledge graph. In *PRCV (1)*, volume 14425 of *Lecture Notes in Computer Science*, pages 395–407. Springer.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023b. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *CoRR*, abs/2307.07697.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023c. Head-to-tail: How knowledgeable are large language models (llm)? A.K.A. will llms replace knowledge graphs? *CoRR*, abs/2308.10168.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020a. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *COLING*, pages 1852–1862. International Committee on Computational Linguistics.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021a. Rpbert: A text-image relation propagation-based BERT model for multimodal NER. In *AAAI*, pages 13860–13868. AAAI Press.

Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. UMIE: unified multimodal information extraction with instruction tuning. *CoRR*, abs/2401.03082.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023d. Generative multimodal models are in-context learners. *CoRR*, abs/2312.13286.

Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020b. Multi-modal knowledge graphs for recommender systems. In *CIKM*, pages 1405–1414. ACM.

Zequn Sun, Muhao Chen, and Wei Hu. 2021b. Knowing the no-match: Entity alignment with dangling cases. In *ACL/IJCNLP (1)*, pages 3582–3593. Association for Computational Linguistics.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *ISWC (1)*, volume 10587 of *Lecture Notes in Computer Science*, pages 628–644. Springer.

Zequn Sun, Wei Hu, Chengming Wang, Yuxin Wang, and Yuzhong Qu. 2023e. Revisiting embedding-based entity alignment: A robust and adaptive method. *IEEE Trans. Knowl. Data Eng.*, 35(8):8461–8475.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020c. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proc. VLDB Endow.*, 13(11):2326–2340.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *ICLR (Poster)*. OpenReview.net.

LLaMA-MoE Team. 2023. Llama-moe: Building mixture-of-experts from llama with continual pre-training.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *CoRR*, abs/2401.06209.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Alasdair Tran, Alexander Patrick Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *CVPR*, pages 13032–13042. Computer Vision Foundation / IEEE.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational AI with personalized social support. *CoRR*, abs/2308.10278.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR (Poster)*. OpenReview.net.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *EMNLP/IJCNLP (1)*, pages 5783–5788. Association for Computational Linguistics.

Fanqi Wan, Xinting Huang, Tao Yang, Xiaojun Quan, Wei Bi, and Shuming Shi. 2023. Explore-instruct: Enhancing domain-specific instruction coverage through active exploration. In *EMNLP*, pages 9435–9454. Association for Computational Linguistics.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021a. Structure-augmented text representation learning for efficient knowledge graph completion. In *WWW*, pages 1737–1748. ACM / IW3C2.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022a. Deepstruct: Pre-training of language models for structure prediction. In *ACL (Findings)*, pages 803–823. Association for Computational Linguistics.

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

Jieming Wang, Ziyan Li, Jianfei Yu, Li Yang, and Rui Xia. 2023a. Fine-grained multimodal named entity recognition and grounding with a generative framework. In *ACM Multimedia*, pages 3934–3943. ACM.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2023b. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *CoRR*, abs/2312.01701.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022c. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. In *ACL (1)*, pages 4281–4294. Association for Computational Linguistics.

Luyao Wang, Pengnian Qi, Xigang Bao, Chunlai Zhou, and Biao Qin. 2024a. Pseudo-label calibration semi-supervised multi-modal entity alignment. In *AAAI*, pages 9116–9124. AAAI Press.

Meng Wang, Yinghui Shi, Han Yang, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. 2023c. Probing the impacts of visual context in multimodal entity alignment. *Data Sci. Eng.*, 8(2):124–134.

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: A large-scale, comprehensive multi-modal knowledge graph. *Big Data Res.*, 22:100159.

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021b. Is visual context really helpful for knowledge graph? A representation learning perspective. In *ACM Multimedia*, pages 2735–2743. ACM.

Peng Wang, Xiaohang Chen, Ziyu Shang, and Wenjun Ke. 2023d. Multimodal named entity recognition with bottleneck fusion and contrastive learning. *IEICE Trans. Inf. Syst.*, 106(4):545–555.

Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022d. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *SIGIR*, pages 938–948. ACM.

Shuo Wang, Meizhi Ju, Yunyan Zhang, Yefeng Zheng, Meng Wang, and Guilin Qi. 2023e. Cross-modal contrastive learning for event extraction. In *DASFAA (3)*, volume 13945 of *Lecture Notes in Computer Science*, pages 699–715. Springer.

Sijia Wang, Alexander Hanbo Li, Henghui Zhu, Sheng Zhang, Pramuditha Perera, Chung-Wei Hang, Jie Ma, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Bing Xiang, and Patrick Ng. 2023f. Benchmarking diverse-modal entity linking with generative models. In *Findings of ACL*, pages 7841–7857. Association for Computational Linguistics.

Xiaodan Wang, Chengyu Wang, Lei Li, Zhixu Li, Ben Chen, Linbo Jin, Jun Huang, Yanghua Xiao, and Ming Gao. 2023g. Fashionklip: Enhancing e-commerce image-text retrieval with fashion multi-modal conceptual knowledge graph. In *ACL (industry)*, pages 149–158. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021c. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Trans. Assoc. Comput. Linguistics*, 9:176–194.

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023h. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *ACM Multimedia*, pages 2391–2399. ACM.

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022e. Named entity and relation extraction with multi-modal retrieval. In *EMNLP (Findings)*, pages 5925–5936. Association for Computational Linguistics.

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022f. Named entity and relation extraction with multi-modal retrieval. In *EMNLP (Findings)*, pages 5925–5936. Association for Computational Linguistics.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022g. ITA: image-text alignments for multi-modal named entity recognition. In *NAACL-HLT*, pages 3176–3189. Association for Computational Linguistics.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022h. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *ACL*, pages 4785–4797. Association for Computational Linguistics.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022i. Promptmner: Prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *DASFAA (3)*, volume 13247 of *Lecture Notes in Computer Science*, pages 297–305. Springer.

Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022j. CAT-MNER: multimodal named entity recognition with knowledge-refined cross-modal attention. In *ICME*, pages 1–6. IEEE.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023i. Self-instruct: Aligning language models with self-generated instructions. In *ACL (1)*, pages 13484–13508. Association for Computational Linguistics.

Yuanyi Wang, Haifeng Sun, Jiabo Wang, Jingyu Wang, Wei Tang, Qi Qi, Shaoling Sun, and Jianxin Liao. 2024b. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. *CoRR*, abs/2401.17859.

Zeqing Wang, Wentao Wan, Runmeng Chen, Qiqing Lao, Minjie Lang, and Keze Wang. 2023j. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *CoRR*, abs/2311.17331.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. 2023k. Learning to filter context for retrieval-augmented generation. *CoRR*, abs/2311.08377.

Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *IJCNN*, pages 1–8. IEEE.

Zilong Wang and Jingbo Shang. 2023. Towards zero-shot relation extraction in web mining: A multimodal approach with relative XML path. In *EMNLP (Findings)*, pages 4254–4265. Association for Computational Linguistics.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023a. Vary: Scaling up the vision vocabulary for large vision-language models. *CoRR*, abs/2312.06109.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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

Haoyang Wen, Ying Lin, Tuan Manh Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Ren Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *NAACL-HLT (Demonstrations)*, pages 133–143. Association for Computational Linguistics.

W. X. Wilcke, Peter Bloem, Victor de Boer, and R. H. van t Veer. 2023. End-to-end learning on multimodal knowledge graphs. *CoRR*, abs/2309.01169.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020a. Scalable zero-shot entity linking with dense entity retrieval. In *EMNLP (1)*, pages 6397–6407. Association for Computational Linguistics.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *CoRR*, abs/2312.14135.

Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023a. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *ACL (1)*, pages 14734–14751. Association for Computational Linguistics.

Tianxing Wu, Chaoyu Gao, Lin Li, and Yuxiang Wang. 2022. Leveraging multi-modal information for cross-lingual entity matching across knowledge graphs. *Applied Sciences*, 12(19):10107.

Yinan Wu, Xiaowei Wu, Junwen Li, Yue Zhang, Haofen Wang, Wen Du, Zhidong He, Jingping Liu, and Tong Ruan. 2023b. Mmpedia: A large-scale multi-modal knowledge graph. In *ISWC*, pages 18–37. Springer.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020b. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *ACM Multimedia*, pages 1038–1046. ACM.

Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How far are we from believable AI agents? A framework for evaluating the believability of human behavior simulation. *CoRR*, abs/2312.17115.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *IJCAI*, pages 3140–3146. ijcai.org.

Xin Xie, Ningyu Zhang, Zhoubo Li, Shumin Deng, Hui Chen, Feiyu Xiong, Mosha Chen, and Huajun Chen. 2022. From discrimination to generation: Knowledge graph completion with generative transformer. In *WWW (Companion Volume)*, pages 162–165. ACM.

Baogui Xu, Chengjin Xu, and Bing Su. 2023a. Cross-modal graph attention network for entity alignment. In *ACM Multimedia*, pages 3715–3723. ACM.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022a. MAF: A general matching and alignment framework for multimodal named entity recognition. In *WSDM*, pages 1215–1223. ACM.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023b. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022b. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *ACM Multimedia*, pages 3857–3866. ACM.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022c. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *ACM Multimedia*, pages 3857–3866. ACM.

Derong Xu, Jingbo Zhou, Tong Xu, Yuan Xia, Ji Liu, Enhong Chen, and Dejing Dou. 2023c. Multimodal biological knowledge graph completion via triple co-attention mechanism. In *ICDE*, pages 3928–3941. IEEE.

Derong Xu, Jingbo Zhou, Tong Xu, Yuan Xia, Ji Liu, Enhong Chen, and Dejing Dou. 2023d. Multimodal biological knowledge graph completion via triple co-attention mechanism. In *ICDE*, pages 3928–3941. IEEE.

Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. 2023e. Urban generative intelligence (UGI): A foundational platform for agents in embodied city environment. *CoRR*, abs/2312.11813.

Guohai Xu, Hehong Chen, Feng-Lin Li, Fu Sun, Yunzhou Shi, Zhixiong Zeng, Wei Zhou, Zhongzhou Zhao, and Ji Zhang. 2021. Alime MKG: A multi-modal knowledge graph for live-streaming e-commerce. In *CIKM*, pages 4808–4812. ACM.

Silei Xu, Shicheng Liu, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu, Sina J. Semnani, and Monica S. Lam. 2023f. Fine-tuned llms know more, hallucinate less with few-shot sequence-to-sequence semantic parsing over wikidata. In *EMNLP*, pages 5778–5791. Association for Computational Linguistics.

Chengmei Yang, Bowei He, Yimeng Wu, Chao Xing, Lianghua He, and Chen Ma. 2023. MMEL: A joint learning framework for multi-mention entity linking. In *UAI*, volume 216 of *Proceedings of Machine Learning Research*, pages 2411–2421. PMLR.

Barry Menglong Yao, Yu Chen, Qifan Wang, Sijia Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. 2023a. AMELI: enhancing multimodal entity linking with fine-grained attributes. *CoRR*, abs/2305.14725.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *CoRR*, abs/1909.03193.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2023b. Exploring large language models for knowledge graph completion. *CoRR*, abs/2308.13916.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *HLT-NAACL*, pages 193–198. The Association for Computational Linguistics.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. 2023. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers. *CoRR*, abs/2401.04695.

Minji Yoon, Jing Yu Koh, Bryan Hooi, and Ruslan Salakhutdinov. 2023. Multimodal graph learning for generative tasks. *CoRR*, abs/2310.07478.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *ACL*, pages 3342–3352. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *ICLR*. OpenReview.net.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *AAAI*, pages 11051–11059. AAAI Press.

Yawen Zeng, Qin Jin, Tengfei Bao, and Wenfeng Li. 2023. Multi-modal knowledge hypergraph for diverse image retrieval. In *AAAI*, pages 3376–3383. AAAI Press.

Zhiwei Zha, Jiaan Wang, Zhixu Li, Xiangru Zhu, Wei Song, and Yanghua Xiao. 2023. M2conceptbase: A fine-grained aligned multi-modal conceptual knowledge base. *CoRR*, abs/2312.10417.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *AAAI*, pages 14347–14355. AAAI Press.

Dongjie Zhang and Longtao Huang. 2022. Multimodal knowledge learning for named entity disambiguation. In *EMNLP (Findings)*, pages 3160–3169. Association for Computational Linguistics.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models.

Jingdan Zhang, Jiaan Wang, Xiaodan Wang, Zhixu Li, and Yanghua Xiao. 2023a. Aspectmmkg: A multi-modal knowledge graph with aspect-aware entities. In *CIKM*, pages 3361–3370. ACM.

Li Zhang, Zhixu Li, and Qiang Yang. 2021b. Attention-based multimodal entity linking with high-quality images. In *DASFAA*, volume 12682 of *Lecture Notes in Computer Science*, pages 533–548. Springer.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Qiang Zhang, Jiazhang Lian, and Huajun Chen. 2022a. Ontoprotein: Protein pretraining with gene ontology embedding. In *ICLR*. OpenReview.net.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. 2023b. Multimodal analogical reasoning over knowledge graphs. In *ICLR*. OpenReview.net.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *AAAI*, pages 5674–5681. AAAI Press.

Qinggang Zhang, Junnan Dong, Hao Chen, Xiao Huang, Daochen Zha, and Zailiang Yu. 2023c. Knowgpt: Black-box knowledge injection for large language models. *CoRR*, abs/2312.06185.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023d. Llmaaa: Making large language models as active annotators. In *EMNLP (Findings)*, pages 13088–13103. Association for Computational Linguistics.

Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023e. Reducing the bias of visual objects in multi-modal named entity recognition. In *WSDM*, pages 958–966. ACM.

Xuan Zhang, Xun Liang, Xiangping Zheng, Bo Wu, and Yuhui Guo. 2022b. MULTIFORM: few-shot knowledge graph completion via multi-modal contexts. In *ECML/PKDD (2)*, volume 13714 of *Lecture Notes in Computer Science*, pages 172–187. Springer.

Yichi Zhang, Mingyang Chen, and Wen Zhang. 2023f. Modality-aware negative sampling for multi-modal knowledge graph embedding. *CoRR*, abs/2304.11618.

Yichi Zhang, Zhuo Chen, Yin Fang, Lei Cheng, Yanxi Lu, Fangming Li, Wen Zhang, and Huajun Chen. 2023g. Knowledgeable preference alignment for llms in domain-specific question answering. *CoRR*, abs/2311.06503.

Yichi Zhang, Zhuo Chen, and Wen Zhang. 2023h. MACO: A modality adversarial and contrastive framework for modality-missing multi-modal knowledge graph completion. In *NLPCC (1)*, volume 14302 of *Lecture Notes in Computer Science*, pages 123–134. Springer.

Yichi Zhang, Zhuo Chen, Wen Zhang, and Huajun Chen. 2023i. Making large language models perform better in knowledge graph completion. *CoRR*, abs/2310.06671.

Yichi Zhang and Wen Zhang. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *CoRR*, abs/2209.07084.

Yuanhan Zhang, Qinghong Sun, Yichun Zhou, Zexin He, Zhenfei Yin, Kun Wang, Lu Sheng, Yu Qiao, Jing Shao, and Ziwei Liu. 2022c. Bamboo: Building mega-scale vision dataset continually with human-machine synergy. *CoRR*, abs/2203.07845.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023j. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Zhengxuan Zhang, Jianying Chen, Xuejie Liu, Weixing Mai, and Qianhua Cai. 2023k. 'what'and 'where'both matter: dual cross-modal graph convolutional networks for multimodal named entity recognition. *International Journal of Machine Learning and Cybernetics*, pages 1–11.

Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. 2022a. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal NER. In *ACM Multimedia*, pages 3983–3992. ACM.

Gang Zhao, Guanting Dong, Yidong Shi, Haolong Yan, Weiran Xu, and Si Li. 2022b. Entity-level interaction via heterogeneous graph for multimodal named entity recognition. In *EMNLP (Findings)*, pages 6345–6350. Association for Computational Linguistics.

Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo. 2021. Boosting entity-aware image captioning with multi-modal knowledge graph. *CoRR*, abs/2107.11970.

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. 2022c. Mose: Modality split and ensemble for multimodal knowledge graph completion. In *EMNLP*, pages 10527–10536. Association for Computational Linguistics.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *ACM Multimedia*, pages 5298–5306. ACM.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021b. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *ICME*, pages 1–6. IEEE.

Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022a. Visual entity linking via multi-modal learning. *Data Intell.*, 4(1):1–19.

Qiushuo Zheng, Hao Wen, Meng Wang, Guilin Qi, and Chaoyu Bai. 2022b. Faster zero-shot multi-modal entity linking via visual-linguistic representation. *Data Intell.*, 4(3):493–508.

Shangfei Zheng, Weiqing Wang, Jianfeng Qu, Hongzhi Yin, Wei Chen, and Lei Zhao. 2023a. MMKGR: multi-hop multi-modal knowledge graph reasoning. In *ICDE*, pages 96–109. IEEE.

Shangfei Zheng, Hongzhi Yin, Tong Chen, Quoc Viet Hung Nguyen, Wei Chen, and Lei Zhao. 2023b. Do as I can, not as I get: Topology-aware multi-hop reasoning on multi-modal knowledge graphs. *CoRR*, abs/2306.10345.

Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. 2021c. Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in bioinformatics*, 22(4):bbaa344.

Wenbo Zheng, Lan Yan, Chao Gou, Zhi-Cheng Zhang, Jun Jason Zhang, Ming Hu, and Fei-Yue Wang. 2021d. Pay attention to doctor-patient dialogues: Multi-modal knowledge graph attention image-text embedding for COVID-19 diagnosis. *Inf. Fusion*, 75:168–185.

Ziqiang Zheng, Yiwei Chen, Jipeng Zhang, Tuan-Anh Vu, Huimin Zeng, Yue Him Wong Tim, and Sai-Kit Yeung. 2024. Exploring boundary of GPT-4V on marine analysis: A preliminary case study. *CoRR*, abs/2401.02147.

Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In *EMNLP*, pages 6293–6302. Association for Computational Linguistics.

Yang Zhou, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Prompting vision language model with knowledge from large language model for knowledge-based VQA. *CoRR*, abs/2308.15851.

Bin Zhu, Meng Wu, Yunpeng Hong, Yi Chen, Bo Xie, Fei Liu, Chenyang Bu, and Weiping Ding. 2023a. MMIEA: multi-modal interaction entity alignment model for knowledge graphs. *Inf. Fusion*, 100:101935.

Bolin Zhu, Xiaoze Liu, Xin Mao, Zhuo Chen, Lingbing Guo, Tao Gui, and Qi Zhang. 2023b. Universal multimodal entity alignment via iteratively fusing modality similarity paths. *CoRR*, abs/2310.05364.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023c. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

Jia Zhu, Changqin Huang, and Pasquale De Meo. 2023d. DFMKE: A dual fusion multi-modal knowledge graph embedding framework for entity alignment. *Inf. Fusion*, 90:111–119.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022a. Multi-modal knowledge graph construction and application: A survey. *CoRR*, abs/2202.05786.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022b. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. 2021. Knowledge perceived multi-modal pretraining in e-commerce. In *ACM Multimedia*, pages 2744–2752. ACM.

A Appendix

A.1 Literature Collection Methodology

For our paper, we source literature primarily from Google Scholar and arXiv. Google Scholar provides broad access to leading computer science conferences and journals, while arXiv serves as a key platform for preprints across various disciplines, including a significant repository recognized by the computer science community. We employ a systematic search strategy on these platforms, using relevant keyword combinations to assemble our references. We rigorously curate this collection, manually filtering out irrelevant papers and incorporating initially overlooked studies mentioned in their main texts. By exploiting Google Scholar's citation tracking, we thoroughly augment our list through iterative depth and breadth traversal.

Organization. § 2 introduces preliminary concepts in KGs and provides an overview of MMKG settings. § 3 reviews the evolution of MMKGs, focusing on the motivations and trends that have shaped their development from inception to their current state. § 4 discusses tasks within the MMKG domain, categorizing them into four key areas: MMKG Acquisition, Fusion, Inference, and MMKG-driven Tasks. This section carefully addresses overlaps across tasks, focusing on core challenges and illustrating them in Fig. 2. Furthermore, § 4.5 analyzes current trends and industrial applications of MMKG, providing insights into their impact across various sectors. Looking ahead, § 5 contemplates the future integration of multimodal methods with MMKGs, proposing potential enhancements for the tasks discussed previously. It also explores opportunities to sustain MMKG growth, especially in light of rapid developments in LLM applications. Finally, § 6 concludes this article.

Related work. Several studies have reviewed literature pertinent to KGs and multi-modal learning. Distinct from these, our survey highlights specific differences.

1) Zhu et al. (2022a) explore various characteristics of mainstream MMKGs and their constructions, primarily from a CV perspective. This include aspects like labeling images with KG symbols and symbol-image grounding. Conversely, Peng et al. (2023) offer a detailed analysis of MMKG from a semantic web perspective, providing a definition and an analysis of its construction and ontology architectures. However, both studies present limited insights into tasks within and beyond MMKG, such as Multi-modal Entity Alignment (MMEA) and Multi-modal Knowledge Graph Completion (MKGC), potentially overlooking MMKG's inherent limitations. To fully grasp the challenges facing MMKG, extensive benchmarks and analyses across various academic and industrial tasks are necessary.
2) Liang et al. (2022) have discussed MMKG reasoning, while Chen et al. (2023d) have explored extraction-based MMKG construction. However, these works, scattered across various tasks, have not been systematically reviewed and analyzed, indicating a need for a cohesive evaluation within the field.
3) The analyses by Zhu et al. (2022a) and Peng et al. (2023) are based on developments up to 2021, whereas the latest discussions by Liang et al. (2022) and Chen et al. (2023d) extend into 2022. This timeline reveals a gap in integrating the most recent insights from the MMKG community. In response to the rapid advancements in AGI from 2022 to 2023, which emphasize emerging areas like LLMs, AI-for-Science, and industrial applications, our survey aims to fill critical knowledge gaps. Our goal is to provide a clear roadmap for future research, highlighting the challenges and opportunities in these fast-evolving fields.

A.2 (MM)KG Preliminaries

Aiming to align with established literature, we begin with a widely-accepted definition of KG and its foundational operations, explore KGs enriched with ontologies from the semantic web perspective.

Multi-modal Learning. We focus on visiolinguistic (VL) tasks involving text and image data, aiming to provide in-depth analysis and research continuity. Other modalities like video or biochemistry are less emphasized as VL methods can often be adapted to them. Thus, the input domain is $\mathcal{X} = \mathcal{X}^{\mathbb{l}} \times \mathcal{X}^{\mathbb{v}}$, with inputs $\hat{x} = (x^{\mathbb{l}}, x^{\mathbb{v}})$, where $x^{\mathbb{l}}$ and $x^{\mathbb{v}}$ are language and visual data, respectively.

A.2.1 Knowledge Graph

Since their inception around 2007, Knowledge Graphs (KGs) have become pivotal in various academic domains, marked by foundational projects such as Yago (Suchanek et al., 2007), DBPedia (Auer et al., 2007), and Freebase (Bollacker et al., 2008). The integration of Google's Knowledge Panels into web search in 2012 highlighted a significant milestone in the adoption of KGs. Today, KGs enhance search engines like Google and Bing and are integral to the functionality of voice assistants like Amazon Alexa and Apple Siri, reflecting their widespread business importance and increasing prevalence.

Definition 1 ***Knowledge Graph.*** *A Knowledge Graph (KG) is denoted as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, comprising an entity set $\mathcal{E}$, a relation set $\mathcal{R}$, and a statement set $\mathcal{T}$. A statement is either a relational fact triple (h, r, t) or an attribute triple (e, a, v). Specifically, KGs consist of a set of relational facts forming a multi-relational graph, wherein nodes represent entities (h and t in $\mathcal{E}$ symbolize head and tail entities, respectively) and edges are denoted by relations ($r \in \mathcal{R}$). Regarding attribute triples, the attribute a ($a \in \mathcal{A}$) indicates that an entity e has a certain attribute with a corresponding value v ($v \in \mathcal{V}$). These values can include various literals, such as strings or dates, and cover metadata like labels and textual definitions, represented through either built-in or custom annotation properties.*

Structural Composition. KGs represent entities and relations using a graph structure, where nodes symbolize real-world entities or atomic values (attributes), and edges denote relations. Knowledge is often captured in triples, such as *(Hangzhou, locatedAt, China)*. They utilize an ontology-based schema to define basic entity classes and their relations, usually in a taxonomic structure. This semi-structured nature merges structured data's clear semantics (from ontologies) with the flexibility of unstructured data, allowing easy expansion through new classes and relations.

Accessibility and Advantages. KGs support a wide array of downstream applications, accessible primarily via *Lookup* and *Querying* methods. *Lookup* in KGs, also known as KG retrieval, identifies relevant entities or properties based on input strings, leveraging lexical indices (surface) from entity and relation labels. An example of this is the DBpedia online lookup service [4]. Alternatively, *Querying* returns results from input queries crafted in the RDF query language SPARQL[5]. These queries typically involve sub-graph patterns with variables, yielding matched entities, properties, literals, or complete sub-graphs.

[4] https://lookup.dbpedia.org/
[5] https://www.w3.org/TR/rdf-sparql-query/

Entity-based KGs Construction. When constructing entity-based KGs, both ontology and data adhere to strict standards, wherein KG nodes typically represent entities in a one-to-one correspondence with real-world objects. These KGs are prominent in both academic projects like Yago and Freebase, and industry initiatives like OpenBG (Dong, 2023) and TeleKG (Chen et al., 2023h).

Note that KGs, especially those with OWL ontologies, support symbolic reasoning, including consistency checks to identify logical conflicts and entailment reasoning to infer hidden knowledge via Description Logics. KGs also facilitate inter-domain connections. An example is the linkage between the *Movie* and *Music* domains through common entities like individuals who are both *actors* and *singers*. This interconnectivity not only enhances machine comprehension but also improves human understanding, benefiting applications like search, question answering, and recommendations. Furthermore, recent developments in LLMs highlight the crucial role of KGs, particularly in managing long-tailed knowledge, as evident in several studies (Dong, 2023; Sun et al., 2023c; Pan et al., 2023a,b).

The construction of these KGs often involves processing entities and relationships from structured sources like relational databases. Wikipedia (Denoyer and Gallinari, 2006), with its entity descriptions and hyperlinks between entity pages, serves as a common starting point for knowledge acquisition. Early KGs like Yago, DBPedia (Auer et al., 2007), and Freebase benefit from the high accuracy of Wikipedia data by transforming Infoboxes into entities and relationships. Additional sources, such as IMDb, MusicBrainz, and Goodreads, enhance coverage, especially for entities of varying popularity.

Integrating knowledge from various structured sources requires tackling three heterogeneity types (Dong, 2023): ***(i)*** **Schema Heterogeneity**, where different data sources may represent the same entity type and relationship differently; ***(ii)*** **Entity Heterogeneity**, where varied source names might depict the same real-world entity; ***(iii)*** **Value Heterogeneity**, where different sources may offer dissimilar or outdated attribute values for identical entities. Addressing these issues has spurred numerous research tasks, including Entity Linking in incomplete KG and data fusion (e.g., KG Completion and Entity Alignment) across diverse KGs. Besides, techniques for extending KG content include extracting knowledge from semi-structured data, such as websites. Here, each page typically represents a topic entity, and information is displayed in key-value pairs, consistently positioned across different pages. These techniques aim to capture long-tail knowledge, often using manually constructed extraction patterns and supervised extraction algorithms.

Text-rich Construction. Unlike entity-based KGs, text-rich KGs, with their dominant text attributes, face challenges in extracting clean, unambiguous entities, making them more akin to bipartite graphs than to conventional connected graphs. Typically, they tolerate greater ambiguities, representing nodes as free texts rather than well-defined entities, making them particularly suited to domains like Products and Encyclopedia where semantic distinctions between values and classes are often unclear (Wang et al., 2021c). The construction of text-rich KGs, especially in domains without a specialized structured knowledge base like Wikipedia, generally depends on extraction models. These models extract structural information from relevant, unstructured source data, employing Named Entity Recognition methods to identify patterns indicative of specific attributes.

A.2.2 Multi-modal Knowledge Graphs

The limitations of traditional uni-modal (text-based) KGs in handling multi-modal applications have driven academic and industrial research to develop Multi-modal Knowledge Graphs (MMKGs). A KG is considered multi-modal (MMKG) when it incorporates knowledge symbols in various modalities, such as text, images, sound, or video. However, in this survey, we primarily focus on the visual modality (i.e., images) beyond traditional text-based KGs.

Specifically, in N-MMKG, a relation triple (h, r, t) in $\mathcal{T}_\mathcal{R}$ may include h or t as an image, with r defining the relation. While in A-MMKG, an attribute triple (e, a, v) in $\mathcal{T}_\mathcal{A}$ might associates an image as v with the attribute a, typically designated as *hasImage*. Note that N-MMKG and A-MMKG are not strictly exclusive: N-MMKG might be considered a particular case of A-MMKG, especially when an entity in A-MMKG takes the form of an image, thereby transforming it into N-MMKG.

Considering that the A-MMKG ontology largely mirrors standard KGs, with the primary distinction being the inclusion of visual attributes, we mainly discuss several representative N-MMKG ontologies in § 3. This emphasis is due to the complex design considerations involved in integrating image entities into N-MMKGs.

MMKGs prior to 2021. Notably, the earliest MMKG in a general sense could be traced back to **ImageNet**(Deng et al., 2009), a large-scale image ontology based on the WordNet (Miller, 1995) structure. Despite its rich semantic hierarchy and millions of annotated images, ImageNet, as an A-MMKG, is primarily utilized for object classification, with its knowledge components often underutilized. **NEIL** (Chen et al., 2013) represents an early effort to construct visual knowledge from the Internet through a cycle of relation extraction, data labeling, and classifiers/detectors learning. However, NEIL's scalability is limited, proved by its intensive computational requirement to classify 400K visual instances of 2273 objects, whereas typical KGs require grounding billions of instances. Further developments (Johnson et al., 2015; Yatskar et al., 2016; Gong and Wang, 2017; Lu et al., 2016) focus on improving visual detection and object segmentation from complex images, with Chen et al. (2014) leveraging learned top-down segmentation priors from visual subcategories to aid in the construction.

Visual Genome (Krishna et al., 2017) provides dense annotations of objects, attributes, and relations, but primarily aids scene understanding tasks like image description and question answering. **ImageGraph** (Oñoro-Rubio et al., 2019), rooted in Freebase (Bollacker et al., 2008), and **IMGpedia** (Ferrada et al., 2017), linking Wikimedia Commons (Commons, 2012) visual data with DBpedia metadata, represents further expansions into MMKGs. ImageGraph, assembled through a web crawler parsing image search results and applying heuristic data cleaning rules (e.g., deduplication and ranking), focuses on reasoning over visual concepts, enabling relation prediction and multi-relational image retrieval. In 2019, Liu et al. (2019b) first formally introduced the term **"MMKG"**, launching three A-MMKG datasets for Link Prediction and Entity Matching research, constructed using a web crawler as the image collector based on Freebase15K (FB15K) (Bordes et al., 2013), averaging 55.8 images per entity. Meanwhile, DBpedia15k (DBP15K) and Yago15k (YG15K) were developed by aligning entities from DBpedia and Yago with FB15K, enriching these KGs with numeric literals, image information, and *sameAs* predicates for cross-KG Entity Linking. **GAIA** (2020) (Li et al., 2020a) is an MMKG extraction system that supports complex graph queries and multimedia information retrieval. It integrates Text Knowledge Extraction and Visual Knowledge Extraction processes on identical document sets, generating modality-specific KGs which are then merged into a coherent MMKG. Concurrently, Then, **VisualSem** (Alberts et al., 2020) emerges as an A-MMKG, sourcing entities and images from BabelNet (Navigli and Ponzetto, 2012) with meticulous filtering to ensure data quality and diversity. Entities in VisualSem are linked to Wikipedia, WordNet synsets (Miller, 1995), and, when available, high-resolution images from ImageNet (Deng et al., 2009). As a N-MMKG, **Richpedia** (Wang et al., 2020) collects images and descriptions from Wikipedia (Vrandecic and Krötzsch, 2014), using hyperlinks and text for manual relationship identification among image entities, supplemented by a web crawler for broader image entity collection.

Recent focus in the MMKG community has shifted from construction to application, emphasizing areas such as MMKG Representation Learning (§ 4.1), Acquisition (§ 4.2), Fusion (§ 4.3), Inference (§ 4.4), and MMKG-driven Applications (§ 4.5). While MMKG acquisition extends construction efforts, it mainly addresses multi-modal extraction challenges (Ma et al., 2022), highlighting the scarcity of large-scale MMKG resources and the demand for task-specific datasets to address MMKG's limitations and support novel downstream tasks. Specifically, Baumgartner et al. (2020) employ multi-modal detectors and a semantic web-informed scheme for semantic relation extraction between movie characters and locations to support Deep Video Understanding.

M^{2}ConceptBase & ManipMob-MMKG. Note that the nodes in M^{2}ConceptBase and Aspect-MMKG are not linked or mapped to existing public KGs. Instead, their focus is on decomposing entity concepts and associating them with fine-grained images. As a result, most nodes within these MMKGs remain isolated, rendering the graphs more akin to multi-modal extensions of text-rich KGs, as discussed in Appendix A.2.1. Song et al. (2023c) unveil a scene-driven MMKG construction method that starts with natural language scene descriptions and employs a prompt-based scene-oriented schema generation. This approach, combined with traditional knowledge engineering and LLMs, streamlines the creation and refinement of the **ManipMob-MMKG**, a specialized MMKG tailored for indoor robotic tasks such as manipulation and mobility.

In-MMKG Task Datasets. Exploring MMKGs' utility in downstream tasks, Xu et al. (2022b) introduce two MMKG Link Prediction datasets, **MKG-W** and **MKG-Y**, derived from OpenEA benchmarks (Sun et al., 2020c) and integrating structured data from Wikipedia/YAGO with expert-validated images sourced from the web. Focusing on Multi-modal Entity Alignment tasks, Li et al. (2023l) introduce **Multi-OpenEA**, extending the OpenEA benchmarks with 16 MMKGs and Google-sourced images. Investigating the effects of the missing visual modality, Chen et al. (2023f) randomly removed images from the DBP15K (Liu et al., 2021) and Multi-OpenEA datasets, releasing the **MMEA-UMVM** datasets. Additionally, Zhang et al. (2023b) define a new task on multi-modal analogical reasoning over KGs, which requires the ability to reason using multiple modalities and background knowledge. They also develop a dataset, MARS, and a corresponding MMKG, **MarKG**, for benchmarking purposes.

N-MMKG Ontology Development. IMGpedia Ontology (Ferrada et al., 2017) (Fig. 3 (a)) extends terms from the DBpedia Ontology and the Open Graph Protocol to represent multi-modal data in RDF. Specifically, the *imo:Image* denotes an abstract resource representing an image, which captures its dimensions (*imo:height*, *imo:width*), URL (*imo:fileURL*), and an *owl:sameAs* link to its corresponding resource in DBpedia Commons. *imo:Descriptor* defines visual descriptors linked via *imo:describes*, with types including *imo:HOG* (Histogram of Oriented Gradient), *imo:CLD* (Color Layout Descriptor), and *imo:GHD* (Gradation Histogram Descriptor). *imo:ImageRelation* encapsulates similarity links between images, detailing the descriptor type used and the Manhattan distance between image descriptors, with an additional *imo:similar* relation for k-nearest neighbor images.

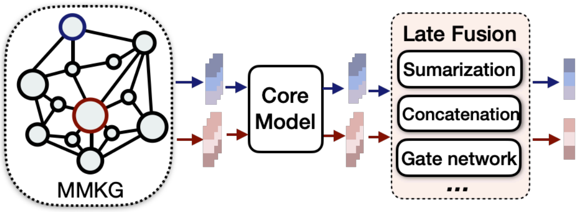
Richpedia ontology (Wang et al., 2020)

(Fig. 3(b)) aligns closely with the IMGpedia Ontology. Here, *rpo:KGEntity* denotes textual KG entities, while *rpo:Image* stands for a Richpedia image entity characterized by a URL and dimensions (e.g., *rpo:Height* and *rpo:Width*, both expressed in the *xsd:float* datatype for numerical values). Subclasses of *rpo:Descriptor*, like *rpo:GHD*, capture visual traits of images. Semantic relations like *rpo:sameAs* and *rpo:imageOf* link these entities, with *rpo:ImageSimilarity* quantifying image likeness between *rpo:sourceImage* and *rpo:targetImage* through pixel-level comparisons. Following Richpedia (Wang et al., 2020), Peng et al. (2023) explore a new MMKG ontology (Fig. 3(c)) to tackle the issue of entities with multiple visual representations (i.e., aspects), a phenomenon emphasized by AspectMMKG (Zhang et al., 2023a) and M^{2}ConceptBase (Zha et al., 2023). The key of this paradigm is to introduce the *Mirror Entity* and *Picture Unit* as foundational concepts. *rpo:MirrorEntity* denotes a particular concept, with *rpo:NamedEntity* pointing to a related KG entity. Its visual counterpart, the *rpo:ImageEntity*, is sourced from the *rpo:PictureUnit*, which might aggregate multiple such image entities under the same aspect. Besides, various *rpo:PictureUnit* maintain a degree of similarity through *rpo:similarity*.An *rpo:align* linkage is established when *rpo:NamedEntity* and *rpo:ImageEntity* both reference a common *rpo:MirrorEntity*. Further, the *rpo:pictureOf* relation binds *rpo:PictureUnit* to *rpo:NamedEntity*, with the *rpo:TextEntity* serving as a bridge, encapsulating shared descriptions. In essence, this ontology enriches the prior MMKG by offering a hierarchical structure, effectively clustering and associating images from diverse aspects.

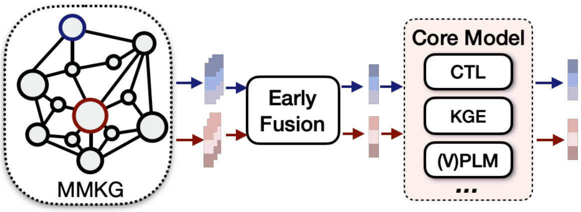
A.3 MMKG Representation Learning

The current mainstream MMKG representation learning approaches primarily concentrate on A-MMKGs, as their similarity to traditional KGs allows for more adaptable paradigm shifts. Those methods for integrating entity modalities within MMKGs generally fall into two categories, which are sometimes overlap within various frameworks, detailed in Fig. 9.

Late Fusion. (Liu et al., 2021; Lin et al., 2022; Li et al., 2022c; Wang et al., 2022b; Lu et al., 2022b). Recent Transformer-based methods (Chen et al., 2023e,f) introduce fine-grained entity-level modality **preference** for entity representation in Multi-modal Entity Alignment.



(a) Late Fusion For MMKG Representation.



(b) Early Fusion For MMKG Representation.

Figure 9: Differences in MMKG representation: Late Fusion focuses on Modality Interaction, applying fusion just before output, while Early Fusion centers on complex reasoning, integrating modalities initially. The former is more oriented towards representation itself, while the latter is more oriented towards cross-modal reasoning. Abbreviations: CTL (Contrastive Learning), KGE (Knowledge Graph Embedding).

Early Fusion. (Fang et al., 2023b; Liang et al., 2023a; Wei et al., 2023b; Chen et al., 2022c; Zhang et al., 2023b) Recent studies (Chen et al., 2022c; Liang et al., 2023a; Zhang et al., 2023b; Lee et al., 2023) utilize (V)PLMs like BERT and ViT for multi-modal data integration.

A.4 MMKG Acquisition

MMKG Acquisition (or Extraction) involves creating an MMKG by integrating multi-modal data such as text, images, audio, and video. This process utilizes multi-modal information from other sources, such as Internet search engines or public databases, either to enhance an existing KG or to develop a new MMKG, thereby enabling a comprehensive understanding of complex, interconnected concepts. The resulting MMKG leverages the unique strengths of each modality to provide a more cohesive and detailed knowledge representation.

A.4.1 Supplementary Information for MNER & MMRE

MNER Definition. MNER is typically considered as a sequence labeling problem, where a model takes a sentence $x^{\mathbb{l}} = \{w_1, w_2, \ldots, w_L\}$ along with an associated image $x^{\mathbb{V}}$ as input to determine the presence and types of named entities in the text. The goal of MNER is to predict a label sequence $\mathcal{Y} = \{y_1, \ldots, y_n\}$, where each label y_i corresponds to a named entity category for each token w_i in the sentence. This process, including the probability calculation for the label sequence, follows foundational sequence labeling techniques in NER (Lample et al., 2016).

As shown in Fig. 5 (left), suppose there is a social media post with a photo of Elon Musk standing in front of a SpaceX signboard, accompanied by a caption: "*Great day at the launch site!*". An **MNER** model would not only use the textual cues ("*Elon Musk*", "*SpaceX*") but also recognize the entities in the image. This visual information reinforces the identification of "*Elon Musk*" as a person and "*SpaceX*" as an organization.

MMRE Definition. MMRE analyzes a sentence $x^{\mathbb{I}} = \{w_1, w_2, \ldots, w_L\}$ alongside a corresponding image $x^{\mathbb{V}}$, focusing on an entity pair (e_1, e_2) within the sentence. The task involves classifying the relationship between these entities, leveraging both textual and visual cues such as object interactions depicted in the image. For each potential relation $r_i \in R$, a confidence score $p(r_i|e_1, e_2, x^{\mathbb{I}}, x^{\mathbb{V}})$ is assigned. The relation set $\mathcal{R} = \{r_1, \ldots, r_C, \texttt{None}\}$ includes pre-defined relation types, with "None" indicating the absence of a specific relation.

As shown in Fig. 5 (right), consider a sports article with a photo of *LeBron James* and *Stephen Curry* during an NBA game, with the caption: "*Epic showdown in tonight's game!*" In this scenario, an **MMRE** model analyzes the text and visual content, interpreting visual cues like their competitive stances and team logos, to infer a opponent and competitive relationship between them as opponents in the game.

Overlap Between MNER & MRE: Typically, both MNER and MMRE enhance text analysis by incorporating visual information, yet they focus on different aspects: MNER on identifying entities, and MMRE on classifying relationships between these entities. In MMKG construction frameworks, MMRE can be considered as a subsequent task to MNER. Despite these differences, the development methods for these tasks are increasingly converging, with many studies employing similar model designs for both MNER and MMRE (Wang et al., 2022e; Chen et al., 2022d; Hu et al., 2023a). Therefore, we discuss them jointly.

MNER Method Details: Advancements in MNER can be marked by diverse approaches to integrating visual and textual information.

- **BiLSTM-based Methods** (Moon et al., 2018b; Lu et al., 2018; Wu et al., 2020b; Sun et al., 2020a; Chen et al., 2021b).
- **PLM-based Methods** (Yu et al., 2020; Wang et al., 2022g, 2023d; Zhang et al., 2021a; Lu et al., 2022a; Xu et al., 2022a; Wang et al., 2022j,f). For example, FMIT (Lu et al., 2022a) leverages flat lattice structure and relative position encoding to enable direct interaction between fine-grained semantic units across different modalities. MAF (Xu et al., 2022a) includes a cross-modal matching module that calculates the similarity score between text and image, using this score to adjust the amount of visual information integrated. Additionally, a cross-modal alignment module aligns the representations of both modalities, creating a unified representation that bridges the semantic gap and facilitates better text-image connections. ITA (Wang et al., 2022g) transforms images into textual object tags and captions for cross-modal input, enabling a text-only PLM to effectively model interactions between modalities and improve robustness against image-related noise. UMGF (Zhang et al., 2021a) leverages graph fusion techniques to effectively combine information from various modalities. Wang et al. (2023d) further propose a Transformer-based bottleneck fusion mechanism that limits noise spread by allowing modalities to interact only through trainable bottleneck tokens. CAT-MNER (Wang et al., 2022j) utilizes entity label-derived saliency scores to refine attention mechanisms, addressing complexities in cross-modal exchanges. MoRe (Wang et al., 2022f) utilizes a multi-modal retrieval framework with distinct textual and image retrievers to gather relevant paragraphs and related images, respectively. This data trains separate models for NER and RE tasks, followed by a Mixture of Experts (MoE) module that synergizes their predictions. TISGF (Cheng et al., 2023a) creates visual and textual scene graphs, encoding them to extract object-level and relationship-level features across modalities. It then employs a text-image similarity module to determine the fusion extent of visual information. Finally, multi-modal features are integrated using a fusion module, with a Conditional Random Fields (CRF) determining

entity types. PromptMNER (Wang et al., 2022i) utilizes entity-related prompts to extract visual clues by assessing their match with an image using the CLIP (Radford et al., 2021). MG-ICL (Guo et al., 2023a) analyzes data at varying granularities, including sentence and word token levels for text, and image and object levels for visuals. Its cross-modal contrast approach enhances text analysis with visual features, supplemented by a visual gate mechanism to filter out noise.

- **Special Cases:** Liu et al. (2023c) propose integrating uncertainty estimation in MNER to improve prediction reliability. Encoder-Decoder-based PLMs like T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), known for their strengths in NLU and NLG, have gained popularity in recent MNER studies. Wang et al. (2023a) introduces a Fine-grained NER and Grounding (FMNERG) task, which involves extracting named entities in text, their detailed types, and corresponding visual objects in images. Here, *(entity, type, object)* triples are converted into a target sequence, and T5 is used to generate this sequence, incorporating a linear transformation layer to adapt the visual object representations into T5's semantic space.

MMRE Method Details: For those PLM-based methods, HVPNet (Chen et al., 2022d) introduces object-level visual information, employing hierarchical visual features and visual prefix-guided fusion for deeper multi-modal integration; DGF-PT (Li et al., 2023g) implements a dual-gated fusion module, using local and global visual gates to filter unhelpful visual data, followed by a generative decoder which leverages entity types to refine candidate relations, thus capturing meaningful visual cues.

- **BiLSTM-based Methods:**
- **PLM-based Methods:**
- **Special Cases:**

Resources & Benchmarks: ***(i)*** **Twitter2015** (Zhang et al., 2018) and **Twitter2017**(Lu et al., 2018): Key MNER datasets featuring diverse multi-modal content from Twitter, covering 2015-2017. They include image-text pairs categorized into Location, Person, Organization, and Miscellaneous. Each record is annotated by experts for named entities. ***(ii)*** **Twitter-FMNERG** (Wang et al., 2023a): Accompanying the Fine-grained NER and Grounding (FMNERG) task, this dataset provides annotations for named entities in text and their corresponding visual objects, including bounding box coordinates. ***(iii)*** **MNRE** (Zheng et al., 2021a): The main dataset for MMRE sourced from Twitter. The brevity of tweets and the varied nature of social media content make MNRE a challenging benchmark for assessing the representation, fusion, and reasoning in multi-modal techniques. ***(iv)*** **JMERE** (Yuan et al., 2023): A joint Multi-modal Entity-Relation Extraction dataset that combines MNER and MMRE.

Table 2: Comparison of MNER performance on the Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018) datasets, evaluated using precision (P), recall (R), and F1 score as metrics. Results for CLIP (Radford et al., 2021) and BLIP (Li et al., 2022a) are sourced from Hu et al. (Hu et al., 2023a).

Models	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
Zhang et al. (2018)	72.75	68.74	70.69	-	-	-
OCSGA (Wu et al., 2020b)	74.71	71.21	72.92	-	-	-
Lu et al. (Lu et al., 2018)	-	-	-	81.62	79.90	80.75
RpBERT (Sun et al., 2021a)	71.15	74.30	72.69	82.85	84.38	83.61
MEGA (Zheng et al., 2021a)	70.35	74.58	72.35	84.03	84.75	84.39
VisualBERT (Li et al., 2019)	68.84	71.39	70.09	84.06	85.39	84.72
IAIK (Chen et al., 2021b)	74.78	71.82	73.27	-	-	-
RIVA (Sun et al., 2020a)	75.02	71.94	73.45	-	-	-
UMT (Yu et al., 2020)	71.67	75.23	73.41	85.28	85.34	85.31
CLIP (Radford et al., 2021)	74.25	74.64	74.44	85.34	85.29	85.31
UMGF (Zhang et al., 2021a)	74.49	75.21	74.85	86.54	84.50	85.51
BFCL (Wang et al., 2023d)	74.02	75.07	74.54	85.99	85.42	85.70
MGCMT (Liu et al., 2024b)	73.57	75.59	74.57	86.03	86.16	86.09
UAMNer (Liu et al., 2022b)	73.02	74.75	73.87	86.17	86.23	86.20
MAF (Xu et al., 2022a)	71.86	75.10	73.42	86.13	86.38	86.25
SMVAE (Zhou et al., 2022)	74.40	75.76	75.07	85.77	86.97	86.37
GEI (Zhao et al., 2022b)	73.39	75.51	74.43	87.50	86.01	86.75
FMIT (Lu et al., 2022a)	75.11	77.43	76.25	87.57	86.26	86.79
DebiasCL (Zhang et al., 2023e)	74.45	76.13	75.28	87.59	86.11	86.84
MRC-MNER (Jia et al., 2022)	78.10	71.45	74.63	88.78	85.00	86.85
HVPNeT (Chen et al., 2022d)	73.87	76.82	75.32	85.84	87.93	86.87
DCM-GCN (Zhang et al., 2023k)	73.41	75.88	74.63	86.09	87.93	87.00
R-GCN (Zhao et al., 2022a)	73.95	76.18	75.00	86.72	87.53	87.11
MPMRC (Bao et al., 2023)	77.15	75.39	76.26	87.10	87.16	87.13
TISGF (?)	71.15	75.35	73.19	86.48	87.78	87.18
MNER-QG (Jia et al., 2023)	77.76	72.31	74.94	88.57	85.96	87.25
MKGformer (Chen et al., 2022c)	-	-	-	86.98	88.01	87.49
DGCF (Mai et al., 2023)	74.76	75.50	75.13	88.50	87.65	88.07
MMIB (Cui et al., 2023b)	74.44	77.68	76.02	87.34	87.86	87.60
ITA (Wang et al., 2022g)	78.93	78.14	78.53	88.52	90.16	89.33
BLIP (Li et al., 2022a)	77.73	76.58	77.15	88.92	88.67	88.79
PromptMNER (Wang et al., 2022i)	78.03	79.17	78.60	89.93	90.60	90.26
CAT-MNER (Wang et al., 2022j)	78.75	78.69	78.72	90.27	90.67	90.47
MoRe (Wang et al., 2022e)	79.33	79.11	79.22	90.74	90.53	90.63
MGICL (Guo et al., 2023a)	80.31	80.06	80.18	91.07	90.61	90.94
PGIM (Li et al., 2023c)	79.21	79.45	79.33	90.86	92.01	91.43
PROMU (Hu et al., 2023a)	80.03	80.97	80.50	91.97	91.33	91.65

A.4.2 Multi-modal Event Extraction

Event Extraction (EE) differs from NER and RE by focusing on the dynamic and temporal aspects of events within data: ***(i)*** **Dynamic Nature**: While NER and RE focus on static aspects of text (i.e., identifying entities and their relationships), EE captures the unfolding and context of events. It in-

Table 3: Comparison of MMRE performance on MNRE (Zheng et al., 2021a).

Models	P	R	F1
MEGA (Zheng et al., 2021a)	64.51	68.44	66.41
MoRe (Wang et al., 2022e)	66.66	70.58	68.56
HVPNet (Chen et al., 2022d)	83.64	80.78	81.85
MKGformer (Chen et al., 2022c)	82.67	81.25	81.95
Wu et al. (Wu et al., 2023a)	84.69	83.38	84.03
DGF-PT (Li et al., 2023g)	84.35	83.83	84.47
Hu et al. (Hu et al., 2023b)	85.03	84.25	84.64
PROMU (Hu et al., 2023a)	84.95	85.76	84.86

volves understanding not just who or what is involved, but also what is happening, when, where, and other event-related details. ***(ii)* Integration of Components**: EE integrates aspects of NER and RE, linking identified entities and their relationships to specific events, thus providing a more complete narrative. ***(iii)* Contextual Richness**: EE delves into the subtleties surrounding event triggers and arguments, offering insights into how events develop and affect the involved entities.

Typically, EE focuses on identifying event **triggers** and **arguments**, capturing the dynamic aspects of events. For example, in the sentence "*The company launched a new product*", "*launched*" is the event trigger, with "*company*" and "*product*" as arguments, indicating the key participants and elements of the event. This concept contrasts with relation and entity in KGs, which primarily represent static entities and their relationships without delving into the evolving nature of events. EE's emphasis on the temporal and contextual aspects of events distinguishes it from the static, entity-focused nature of KGs, highlighting its unique role in dynamic data analysis and knowledge representation.

Early text-based EE methods leverage techniques like CNNs (Nguyen and Grishman, 2015) and RNNs (Nguyen et al., 2016; Liu et al., 2019a, 2020), with subsequent models adopting GNNs (Li et al., 2017) to better understand event-context dependencies. The advent of PLMs further improve EE capabilities (Wadden et al., 2019; Wang et al., 2022a; Lu et al., 2022c). In CV field, EE aligns with situation recognition (Pratt et al., 2020; Khan et al., 2022), focusing on identifying visual events in images or videos. This progression reflects a broader shift towards a more holistic understanding of events in diverse contexts, paving the way for the development of Multi-modal Event Extraction (MMEE).

Definition 2 ***Multi-modal Event Extraction.*** *MMEE simultaneously analyze textual data (e.g., sentences or paragraphs) $x^{\mathbb{I}} = \{w_1, w_2, ..., w_n\}$ and visual data (e.g., images or videos) $x^{\mathbb{V}}$, both potentially annotated with predefined event types $\mathcal{Y}_e$ and argument types $\mathcal{Y}_a$. In a multi-modal document $\mathcal{D} = \{\mathcal{X}^{\mathbb{I}}, \mathcal{X}^{\mathbb{V}}\}$, an event mention m is classified under an event type y_e and is identified by a trigger, which can be a word w, an image $x^{\mathbb{V}}$, or both. The task extends to extracting and classifying all event participants (i.e., arguments) within $\mathcal{D}$, assigning each to a specific argument type y_a. Arguments are based on textual spans or object bounding boxes in the image, with their positions explicitly identified.*

Methods: Some works (Li et al., 2020b; Chen et al., 2021a; Du et al., 2023b) focus on region feature refinement for MMEE. Specifically, WASE (Li et al., 2020b) utilizes graphical representations of multi-modal documents for cross-modal event co-reference and image-sentence matching, targeting the challenge of limited multi-modal event annotations with a weakly supervised approach which leverages annotated uni-modal corpora and an image-caption alignment dataset. JMMT (Chen et al., 2021a) employs multi-instance learning to assess region and sentence combinations, identifying key areas for multi-modal event co-reference and linking events across visual and textual modalities. CAMEL (Du et al., 2023b) enhances object representation in images by focusing on three specific areas within each object's bounding box and averages the encoded embeddings to aid argument extraction.

Recent advances emphasize refining representations via Contrastive Learning (CL) (Li et al., 2022b; Wang et al., 2023e; Li et al., 2023a). Concretely, CLIP-EVENT (Li et al., 2022b) contrasts images with event-aware text descriptions to training the VLMs; CoCoEE (Wang et al., 2023e) employs CL with weighted samples according to event frequency; TSEE (Li et al., 2023a) aligns optical flow with event triggers and types, observing a strong correlation between similar motion patterns and identical triggers with multi-level CL.

Moreover, emerging research explores zero-shot (Liu et al., 2022a) and few-shot (Moghimifar et al., 2023) approaches to MMEE, potentially enhancing model adaptability to new or sparse data scenarios.

Resources & Benchmarks: ***(i)* M2E2** (Li et al., 2020b): Comprising multi-media news articles

Table 4: Comparative analysis of MMEE results across diverse datasets. M2E2 (Li et al., 2020b) utilizes image and text inputs. Both TVEE (Chen et al., 2021a) and VM2E2 (Wang et al., 2023e) employ video and text inputs.

Dataset	Models	Trigger			Argument		
		P	R	F1	P	R	F1
M2E2	Flat (Li et al., 2020b)	33.9	59.8	42.2	12.9	17.6	14.9
	WASE (Li et al., 2020b)	38.2	67.1	49.1	18.6	21.6	19.9
	CLIP-EVENT (Li et al., 2022b)	41.3	72.8	52.7	21.1	13.1	17.1
	UniCL (Liu et al., 2022a)	44.1	67.7	53.4	24.3	22.6	23.4
	CAMEL (Du et al., 2023b)	55.6	59.5	57.5	31.4	35.1	33.2
TVEE	JMMT (Chen et al., 2021a)	74.3	80.2	77.1	50.1	54.9	52.3
	CoCoEE (Wang et al., 2023e)	80.7	76.4	78.5	65.6	45.4	53.6
	TSEE (Li et al., 2023a)	82.6	80.5	81.5	67.0	49.3	56.8
VM2E2	JMMT (Chen et al., 2021a)	39.7	56.3	46.6	17.9	24.3	20.6
	CoCoEE (Wang et al., 2023e)	47.3	47.7	47.5	26.7	18.5	21.8
	TSEE (Li et al., 2023a)	49.2	53.5	51.6	24.5	27.4	25.9

from the Voice of America website (2016-2017), M2E2 covers a wide range of topics like military affairs, economy, and health. ***(ii)*** **VOANews** (Li et al., 2022b): Constructed with image captions from various news websites, selected for their event-rich content, VOANews aims to provide a challenging benchmark for image retrieval tasks. ***(iii)*** **VM2E2** (Chen et al., 2021a): This first text-video dataset for MMEE is curated using YouTube searches with event types and news source names, focusing on sources like VOA, BBC, and Reuters. ***(iv)*** **TVEE** (Wang et al., 2023e): TVEE features international news videos with captions from the On Demand News channel, aligning with the ACE2005 benchmark's partial event types.

Metrics: Precision (P), recall (R), and F1 score are pivotal in evaluating these tasks. Precision is the ratio of correctly identified entities (or relations) to the total identified. E.g., in MNER, it reflects the proportion of accurately identified named entities from text and associated multi-modal data. Recall is the ratio of correctly identified entities (or relations) to the total relevant entities (or relations) in the dataset. E.g., in MMEE, it gauges the accuracy of extracting entities from text and multi-modal content. The F1 score, harmonizing precision, and recall, offers a comprehensive measure of both metrics. E.g., in MMRE, it provides an equilibrium, assessing the system's performance in discerning text-based entity relationships, integrating precision and recall considerations.

Discussion 1 *Recent advancements for these tasks show a trend towards unified model designs, as evidenced by a range of studies (Wang et al., 2022e; Chen et al., 2022d; Hu et al., 2023a; Cui et al., 2023a; Sun et al., 2024). In certain MMEE datasets such as VM2E2 (Chen et al., 2021a), the visual modality lacks direct event and argument annotations, positioning visual features as supportive elements in benchmarking. However, the prevalent multi-modal F1 score, focusing mainly on text-based event type classification, overlooks the contribution evaluation of visual elements. This scenario highlights the need for future research to devise more balanced multi-modal evaluation metrics that thoroughly integrate visual and textual components. Looking forward, the emergence of MLLMs and their zero-shot extraction capabilities (Wei et al., 2022; Li et al., 2023d) heralds a pivot towards generative-based approaches. This shift implies a broader horizon for MNER, MMRE, and MMEE, urging the expansion into more intricate, specialized, and inherently comprehensive multi-modal extraction tasks.*

A.5 MMKG Fusion

This process involves various tasks, including Multi-Modal Entity Alignment (MMEA), Entity Linking (MMEL), and Entity Disambiguation (MMED).

A.5.1 Supplementary Information for MMEA

A MMKG is denoted as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{T}, \mathcal{V}\}$ with $\mathcal{T} = \{\mathcal{T}_\mathcal{A}, \mathcal{T}_\mathcal{R}\}$. Given two aligned **A-MMKGs** $\mathcal{G}_1 = \{\mathcal{E}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{V}_1, \mathcal{T}_1\}$ and $\mathcal{G}_2 = \{\mathcal{E}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{V}_2, \mathcal{T}_2\}$, the goal of MMEA is to identify pairs of entities (e_i^1, e_i^2) from $\mathcal{E}_1$ and $\mathcal{E}_2$ respectively, that represent the same real-world entity e_i. A set of pre-aligned entity pairs serves as a reference, divided into a training set (seed alignments $\mathcal{S}$) and a test set $\mathcal{S}_{te}$, proportioned by a pre-defined seed alignment ratio R_{sa}. The available modalities associated with an entity are denoted by $\mathcal{M} = \{\mathbb{g}, \mathbb{r}, \mathbb{a}, \mathbb{v}, \mathbb{s}\}$, which represent the graph structure, relation, attribute, vision, and surface (i.e., entity name) modalities, respectively.

Traditional Entity Alignment (EA). Specifically, symbolic logic approaches (Qi et al., 2021) apply manually defined rules, such as logical inference and lexical matching, to guide the alignment. Embedding-based methods (Sun et al., 2023e) utilize learned entity embeddings to expedite the alignment, without predefined heuristics.

MMEA Considerations. While both relation, attribute, and surface modalities can be categorized under language modalities, they are frequently distinguished as separate modalities in MMEA communities (Liu et al., 2021; Lin et al., 2022; Cheng et al., 2022; Chen et al., 2023e,f; Guo et al., 2023b; Su et al., 2023; Zhu et al., 2023d). Besides, research shows a variety of modal usage patterns: some studies focus solely on the **types** of attributes and relations during the alignment process (Chen et al., 2023e,f), while others incorporate their **textual content** into entity representations via using PLM (e.g., BERT (Devlin et al., 2019)) (Wu et al., 2022; Zhu et al., 2023a,b; Li et al., 2023i; Ge et al., 2021; Congcong Ge and Xiaoze Liu and Lu Chen and Baihua Zheng and Yunjun Gao, 2021) or word embeddings (e.g., Glove (Pennington et al., 2014)) (Liu et al., 2021; Lin et al., 2022; Chen et al., 2023e,f, 2022b). Additionally, some methods are proposed for entities that have only one image (Liu et al., 2021; Lin et al., 2022), while others are prepared to handle cases where the number of images per entity can be multiple (Li et al., 2023l) or even missing (Chen et al., 2023f).

MMEA Method Details:

- **Exploring better cross-KG modality feature fusion:** Specifically, MMEA (Chen et al., 2020) is first introduced in 2020 as a method that merges knowledge representations from multiple modalities and aligns entities by minimizing the distance between their holistic embeddings; HMEA (Guo et al., 2021) expands MMKG representation from the Euclidean space to the hyperbolic manifold, offering a more refined geometric interpretation. EVA (Liu et al., 2021) assigns different importance to each modality via an attention mechanism. It further introduces an unsupervised MMEA approach that leverages visual similarities between entities to create a pseudo seed dictionary, thus reducing dependence on gold-standard labels. MSNEA (Chen et al., 2022b) leverages visual cues to guide relational feature learning and weights valuable attributes for alignment. MCLEA (Lin et al., 2022) applies KL divergence to bridge the modality distribution gap between joint and uni-modal embedding. ACK-MMEA (Li et al., 2023h) presents an attribute-consistent KG representation learning method to solve the contextual gap caused by different attributes. PathFusion (Zhu et al., 2023b) combines information from different modalities using the modality similarity path as an information carrier. DFMKE (Zhu et al., 2023d) employs a late fusion approach with modality-specific low-rank factors that enhance feature integration across various knowledge spaces, complementing early fusion output vectors. Considering that the surrounding modality of each entity is inconsistent, MEAformer (Chen et al., 2023e) dynamically adjusts the mutual modality preference for entity-level modality fusion. Recent works like MoAlign (Li et al., 2023i), UMAEA (Chen et al., 2023f) PCMEA (Wang et al., 2024a) and DESAlign (Wang et al., 2024b) follow similar settings. XGEA (Xu et al., 2023a) leverages the information from one modality as complementary relation information to enrich entity embeddings by computing inter-modal attention within the GAT layers.
- **Analyzing the practical limitations and challenges in MMKG alignment:** Wang et al. (2023c) tackled the issue of image-type mismatches in aligned multi-modal entities by filtering out incongruent images using predefined ontologies and an image type classifier. The inherent incompleteness of visual data in MMKGs poses another challenge, where many entities lack images (e.g., 67.58% in DBP15K$_{JA\text{-}EN}$ (Liu et al., 2021)). Furthermore, the intrinsic ambiguity of visual images also impacts the alignment quality (i.e., each entity has multiple visual aspects as elaborated in § 2). Chen et al. (2023f) introduces the MMEA-UMVM dataset to study the impact of training noise and performance degradation at high rates of missing modalities. They further propose UMAEA, which employs a multi-scale modality hybrid approach with a circularly missing modality imagination module equipped. Considering that many entities in the source KG may not have aligned entities in the target KG (i.e., the dangling entities (Sun et al., 2021b; Luo and Yu, 2022)), Guo et al. (2023b) introduce the entity synthesis task to generate new entities either conditionally or unconditionally, and propose the GEEA framework, which employs a mutual variational autoencoder (M-VAE) for entity synthesis. To overcome the costly and time-intensive process of acquiring initial seeds, Ni et al. (2023) developed the Pseudo-Siamese Network (PSNEA), complemented by an Incremental Alignment Pool that labels probable alignments, reducing reliance on data swapping and sample re-weighting.

Discussion 2 *Adopting strategies beyond model ar-*

chitecture is recognized for boosting performance. Iterative training (Lin et al., 2022; Liu et al., 2021), for example, incrementally refines model performance by identifying and adding cross-KG entity pairs as mutual nearest neighbors in the embedding space every K_e epochs (e.g., 5), with pairs confirmed for inclusion in the training set after remaining mutual nearest neighbors across K_s successive iterations (e.g., 10). Similarly, the STEA framework (Liu et al., 2023a) can be utilized to generate additional pseudo-aligned pairs, thereby expanding the training data. Additionally, the CMMI module (Chen et al., 2023f) can be integrated into models to create synthetic visual embeddings, mitigating the impact of missing images. For fair evaluation, models employing these strategies should be assessed separately from those that do not. Moreover, considerations like the use of entity names (surface forms), computational complexity, textual encoding methods, and the integration of additional data warrant careful attention in comparing methodologies in future research.

Resources & Benchmarks: ***(i)*** The first MMEA dataset includes FB15K-DB15K (**FBDB15K**) and FB15K-YAGO15K (**FBYG15K**) (Liu et al., 2019b) with three data splits: $R_{sa} \in \{0.2, 0.5, 0.8\}$. ***(ii)*** **Multi-modal DBP15K** (Liu et al., 2021): An extension of the DBP15K (Sun et al., 2017) which attaches entity-matched images from DBpedia (Auer et al., 2007) and Wikipedia (Denoyer and Gallinari, 2006) to the original cross-lingual EA benchmark. It includes four language-specific KGs from DBpedia, with three bilingual settings ($R_{sa} = 0.3$), namely DBP15K$_{ZH\text{-}EN}$, DBP15K$_{JA\text{-}EN}$, and DBP15K$_{FR\text{-}EN}$. Each setting contains approximately 400K triples and 15K pre-aligned entity pairs. We benchmark those recent MMEA methods using this series of datasets as outlined in Table 5. ***(iii)*** **Multi-OpenEA** (Li et al., 2023l): A multi-modal expansion of the OpenEA benchmarks (Sun et al., 2020c) which links entities with their top-3 related images sourced through Google search. ***(iv)*** **MMEA-UMVM**(Chen et al., 2023f): It contains two bilingual datasets (EN-FR-15K, EN-DE-15K) and two monolingual datasets (D-W-15K-V1, D-W-15K-V2) derived from Multi-OpenEA datasets ($R_{sa} = 0.2$) (Li et al., 2023l) and all three bilingual datasets from DBP15K (Liu et al., 2021). It introduces variability in visual information by randomly removing images, resulting in 97 distinct dataset splits.

Table 5: Comparison of MMEA results with (w/o) and without (w/o) surface forms (SF) on the DBP15K dataset (Liu et al., 2021), where "iter." signifies iterative learning applied. The symbol † indicates that the PLMs were applied for generating surface or attribute embeddings. $*$ marks the results reproduced in (Chen et al., 2023f,e; Xu et al., 2023a).

	Models	DBP15K$_{ZH\text{-}EN}$		DBP15K$_{JA\text{-}EN}$		DBP15K$_{FR\text{-}EN}$	
		H@1	MRR	H@1	MRR	H@1	MRR
w/o SF	HMEA (Guo et al., 2021)	.540	-	.531	-	.484	-
	EVA (Liu et al., 2021)	.720	.793	.716	.792	.715	.795
	MCLEA* (Lin et al., 2022)	.726	.796	.719	.789	.719	.792
	GEEA (Guo et al., 2023b)	.761	.827	.755	.827	.776	.844
	MEAformer (Chen et al., 2023e)	.772	.835	.769	.840	.771	.841
	UMAEA (Chen et al., 2023f)	.800	.860	.801	.862	.818	.877
	DESAlign (Wang et al., 2024b)	.810	.865	.811	.869	.826	.885
w/o SF (iter.)	EVA (Liu et al., 2021)	.761	.814	.762	.817	.793	.847
	MSNEA* (Chen et al., 2022b)	.821	.877	.805	.849	.822	.859
	PSNEA (Ni et al., 2023)	.811	.858	.807	.846	.843	.871
	MCLEA (Lin et al., 2022)	.816	.865	.812	.865	.834	.885
	MEAformer (Chen et al., 2023e)	.847	.892	.842	.892	.845	.894
	SKEA (Su et al., 2023)	.849	.897	.844	.895	.878	.921
	UMAEA (Chen et al., 2023f)	.856	.900	.857	.904	.873	.917
	DESAlign (Wang et al., 2024b)	.868	.909	.871	.913	.882	.924
	XGEA (Xu et al., 2023a)	.876	.910	.878	.914	.889	.924
w/ SF	CLEM† (Wu et al., 2022)	.854	.879	.885	.904	.936	.952
	MSNEA* (Chen et al., 2022b)	.887	.913	.938	.955	.969	.980
	EVA* (Liu et al., 2021)	.929	.951	.964	.976	.990	.994
	MCLEA* (Lin et al., 2022)	.926	.946	.961	.973	.987	.992
	MEAformer (Chen et al., 2023e)	.949	.965	.978	.986	.991	.995
w/ SF (iter.)	MSNEA* (Chen et al., 2022b)	.896	.922	.942	.958	.971	.982
	EVA (Liu et al., 2021)	.956	.969	.979	.987	.995	.997
	SKEA (Su et al., 2023)	.913	.938	.923	.948	.978	.985
	MCLEA (Lin et al., 2022)	.972	.981	.986	.991	.997	.998
	XGEA (Xu et al., 2023a)	.968	.978	.985	.991	.994	.996
	MEAformer (Chen et al., 2023e)	.973	.983	.991	.995	.996	.998

A.5.2 Multi-modal Entity Linking

Entity Linking (EL) serves as a crucial component in various applications (Shen et al., 2014, 2021; Sevgili et al., 2022), including Question Answering, Relation Extraction, and Semantic Search. The main target of EL is to associate textual mentions within documents with their respective entities in a KG (e.g., Freebase (Bollacker et al., 2008)). Notably, mentions extend beyond textual forms, including images, audio, and video content, all of which can be linked to KG entities. Recent studies in Multi-Modal Entity Linking (MMEL) find that leveraging the multi-modal information can significantly enhance the efficacy of conventional EL methods.

Definition 3 ***Multi-modal Entity Linking.*** *A MMKG is denoted as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{T}, \mathcal{V}\}$, where $\mathcal{E} = \{e_1, e_2, ..., e_i\}$ are the entity set. $\mathcal{M} = \{\mathbb{g}, \mathbb{r}, \mathbb{a}, \mathbb{v}, \mathbb{s}\}$ are the graph structure, relation, attribute, vision, and surface information, respectively. For example, $x^{\mathbb{s}}_{e_1}$, $x^{\mathbb{v}}_{e_1}$ denotes the name and visual information of e_1, respectively. The* *mention set is defined as* $\mathcal{N} = \{m_1, ..., m_i\}$ *with* $\{x^{\mathbb{S}}_{m_1}, ..., x^{\mathbb{S}}_{m_i}\}$, $\{x^{\mathbb{V}}_{m_1}, ..., x^{\mathbb{V}}_{m_i}\}$ *being the corresponding name and visual information. The objective of MMEL is to determine the linkage between entities and mentions, denoted by* (e_i, m_i)*, based on the multi-modal information* $(x^{\mathbb{S}}_{e_1}, ..., x^{\mathbb{V}}_{e_1}, x^{\mathbb{S}}_{m_1}, ..., x^{\mathbb{V}}_{m_1})$.

Method: Early MMEL research (Moon et al., 2018a; Adjali et al., 2020; Zhang et al., 2021b) focuses on fusing and expanding multi-modal data, such as merging visual and textual elements from media posts, to enhance textual mentions and predict corresponding KB entities. For example, DZMNED (Moon et al., 2018a) utilizes KG embeddings along with a blend of word-level and char-level lexical embeddings, a strategy crafted to adeptly manage the challenge of identifying previously unseen entities during testing. Zhang et al. (2021b) focus on the removal of noisy images to enhance performance. Subsequent research extends these methods, exploring strategies for integrating diverse multi-modal contexts and developing more reasonable multi-modal datasets (Gan et al., 2021; Zheng et al., 2022a,b; Wang et al., 2022d; Yang et al., 2023; Luo et al., 2023; Yao et al., 2023a). GHMFC (Wang et al., 2022d), for example, employs gated fusion and contrastive training for improved mention representations, while MIMIC (Luo et al., 2023) introduces a multi-grained interaction network for universal feature extraction. AMELI (Yao et al., 2023a) implements an entity candidate retrieval pipeline, enhancing MMEL models using attribute information.

Recent explorations in MMEL mainly employ (V)PLMs for feature representation. BERT (Devlin et al., 2019) is frequently used for textual processing (Yang et al., 2023; Wang et al., 2023f), while CLIP (Radford et al., 2021) is preferred for visual encoding (Song et al., 2023b; Shi et al., 2023). Typically, most parameters of these (V)PLMs remain frozen, complemented by focused fine-tuning strategies. Among them, GEMEL (Shi et al., 2023) effectively combines LLaMA (Touvron et al., 2023) for language processing and CLIP for visual encoding, showing the potential of GPT 3.5 in MMEL. Yang et al. (2023) introduce a multi-mention MMEL task that considers different mentions within the same context as a single sample, employing a multi-mention collaborative ranking method for testing to uncover potential connections between mentions. Pan et al. (2022a) present Multi-modal Item-aspect Linking, focusing on linking short videos to related items in a short-video encyclopedia. GDMM (Wang et al., 2023f) approaches MMEL by incorporating all three modalities: text, image, and table, adhering to a multi-modal encoder-decoder paradigm. DWE (Song et al., 2023b) augments visual features with detailed image attributes, like facial characteristics and scene features, enhancing textual representations using Wikipedia descriptions which bridges the gap between text and KG entities.

Resources & Benchmarks: ***(i)*** **SnapCaptionsKB** (Moon et al., 2018a): A MMEL dataset featuring 12,000 manually labeled image-caption pairs, designed to capture diverse multi-modal interactions. Currently unavailable due to the General Data Protection Regulation (GDPR). In response, Adjali et al. (2020) develop an automated MMEL dataset construction tool from Twitter. ***(ii)*** **M3EL** (Gan et al., 2021): A dataset comprising 181,240 textual mentions and 45,297 images related to movies, offering fine-grained annotations. ***(iii)*** **NYTimes-MEL** (Yang et al., 2023): Originates from the New York Times' (Tran et al., 2020; Zhao et al., 2021) images and captions, focusing on `PERSON` entities. StanfordNLP tool (Qi et al., 2018) is used for NER in captions, where some entities were replaced with nicknames for mention construction. Similar to (Wang et al., 2022d), it is enriched with images and 14 properties for each entity from Wikidata (Xu et al., 2023f), excluding samples with invalid entities or those without corresponding images. ***(iv)*** **WikiData-Based Datasets**: Including **WikiDiverse** (Wang et al., 2022h) and **WikiMEL** (Wang et al., 2022d), these datasets offer human-annotated mentions spanning diverse topics and entity types. WikiDiverse includes data from WikiNews categories like sports and technology, while WikiMEL collates mentions from Wikipedia and WikiData.

Discussion 3 *Evaluation metrics commonly used in this field include Hits@k (e.g., Hits@1, 3, 5), MRR, and MR. These metrics necessitate calculating the similarity or probability between a mention and all entities in the KG. Typically, encoders' parameters are not trained from scratch; instead, employing existing LLMs and vision encoders is standard practice. While many methods permit gradient updates for these parameters, recent findings suggest that maintaining them in a frozen state can markedly decrease training costs while still achiev-*

Table 6: Comparison of MMEL results on the WikiMEL (Wang et al., 2022d) and Wikidiverse (Wang et al., 2022h) dataset.

	Models	WikiMEL			Wikidiverse		
		H@1	H@5	MRR	H@1	H@5	MRR
Text	BLINK (Wu et al., 2020a)	.747	.906	.817	.571	.853	.692
	BERT (Devlin et al., 2019)	.748	.905	.818	.558	.831	.674
	RoBERTa (Liu et al., 2019c)	.738	.898	.809	.595	.851	.705
	GENRE (Cao et al., 2021)	.601	-	-	.601	-	-
	GPT 3.5 Turbo	.727	-	-	.738	-	-
Text + Vision	JMEL (Adjali et al., 2020)	.647	.834	.734	.374	.610	.482
	DZMNED (Moon et al., 2018a)	.788	.926	.850	.569	.814	.676
	GHMFC (Wang et al., 2022d)	.765	.920	.834	.603	.847	.710
	CLIP (Radford et al., 2021)	.832	.945	.882	.612	.852	.717
	ViLT (Kim et al., 2021)	.726	.879	.795	.344	.578	.452
	MMEL (Yang et al., 2023)	.715	.917	-	-	-	-
	GEMEL (Shi et al., 2023)	.826	-	-	.863	-	-
	ALBEF (Li et al., 2021)	.786	.918	.846	.606	.813	.699
	METER (Dou et al., 2022)	.725	.882	.795	.531	.776	.637
	MIMIC (Luo et al., 2023)	.880	.964	.918	.635	.864	.734

ing, or even surpassing, competitive performance levels.

A.5.3 Multi-modal Entity Disambiguation

In many studies, EL and Entity Disambiguation (ED) are often treated synonymously due to their methodological and task-setting similarities (Moon et al., 2018a; Luo et al., 2023). However, it is crucial to distinguish between the two. While EL includes the broader process of identifying and linking named entities in text to their corresponding entities in a KG, ED specifically focuses on resolving cases where a named entity might correspond to multiple potential candidates. In ED, each dataset sample typically includes a named entity alongside a set of candidates that bear close resemblance, highlighting the task's emphasis on disambiguating among these options (Moon et al., 2018a).

Although EL and Entity Disambiguation (ED) are often treated synonymously in many studies due to their methodological and task-setting parallels (Moon et al., 2018a; Luo et al., 2023), distinguishing between them is still vital. EL includes the broader process of identifying and linking named entities in text to their corresponding entries in a KG. In contrast, ED specifically targets resolving ambiguities when a named entity could match multiple candidates. ED emphasizes disambiguating among these potential candidates, often presented with a named entity and a closely related set of options in each dataset sample.

In Multi-modal Entity Disambiguation (MMED), methods leverage not just textual but also visual information to refine disambiguation. For example, DZMNED (Moon et al., 2018a) utilizes a convolutional LSTM for integrating multi-modal data. ET (Adjali et al., 2020) applies an Extra-Tree Classifier to effectively distinguish among ambiguous candidates. IMN (Zhang and Huang, 2022) adopts meta-learning for multi-modal knowledge acquisition and a knowledge-guided transfer learning strategy, facilitating the extraction of cohesive representations across modalities.

A.6 MMKG Inference

This stage, following extraction and fusion within the MMKG construction cycle, aims to bolster the model's reasoning abilities and deepen its understanding of the KG's overall knowledge.

A.7 Supplementary Information for MKGC

Multi-modal Knowledge Graph Completion (MKGC) plays a vital role in mining missing triples from existing KGs. This process involves three sub-tasks: Entity Prediction, Relation Prediction, and Triple Classification.

Definition 4 ***MMKG Completion.*** *A MMKG is denoted as $\mathcal{G} = \mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{T}, \mathcal{V}$, where $\mathcal{T} = \mathcal{T}_{\mathcal{A}}, \mathcal{T}_{\mathcal{R}}$. The goal of MKGC is to enrich the relational triple set $\mathcal{T}_R$ in* ***A-MMKGs*** *by identifying missing relational triples among existing entities and relations, potentially leveraging attribute triples $\mathcal{T}_A$. Specifically, Entity Prediction determines missing head/-tail entities in queries $(h, r, ?)$ or $(?, r, t)$; Relation Prediction identifies missing relations in $(h, ?, t)$; and Triple Classification assesses the validity of given triples (h, r, t) as true or false.*

Methods: Mainstream MKGC approaches primarily follow two paths: embedding-based and fine-tuning (FT) based methods. Considering the intersection between MKGC and KGC methods, this section also discusses several typical KGC techniques to offer deeper insights into MKGC.

Embedding-based Approaches evolve from traditional KGE techniques (Bordes et al., 2013; Sun et al., 2019), adapting them to include multi-modal data, thus forming multi-modal entity embeddings. They're divided into modal fusion, modal ensemble, and negative sampling approaches:

(i) **Modality Fusion** methods (Wilcke et al., 2023; Wang et al., 2022b; Huang et al., 2022) integrate multi-modal embeddings of entities with their structural embeddings for triple plausibility estimation. Early efforts, like IKRL (Xie et al., 2017), use multiple TransE-based scoring functions (Bordes et al., 2013) for modal interaction. Subsequent developments, like TBKGC (Sergieh et al., 2018), TransAE (Wang et al., 2019), and MKBE (Pezeshkpour et al., 2018) further incorporate modalities such as textual numerical attributes. RSME (Wang et al., 2021b) introduces gates for adaptive modal information selection. OTKGE (Cao et al., 2022b) applies optimal transport for multi-modal fusion, while CMGNN (Fang et al., 2023a) implements a multi-modal GNN with cross-modal contrastive learning. HRGAT (Liang et al., 2023b) builds a hyper-node relational graph for multi-modal entity representation. CamE (Xu et al., 2023c) introduces a triple co-attention module for biological KGs, and VISITA (Lee et al., 2023) develops a transformer-based framework which utilizes relation and triple-level multi-modal information for MKGC.

***(ii)* Modality Ensemble** methods train separate models using distinct modalities, combining their outputs for final predictions. For example, MoSE (Zhao et al., 2022c) utilizes structural, textual, and visual data to train three KGC models, using ensemble strategies for joint predictions. Similarly, IMF (Li et al., 2023k) proposes an interactive model to achieve modal disentanglement and entanglement to make robust predictions.

***(iii)* Modality-aware Negative Sampling** involves generating false triples to enhance a model's ability to differentiate between accurate and potentially erroneous KG triples. During KG Embedding training, models map entities and relations to vectors, guided by both positive and negative samples, with their efficacy relying on the strategic selection and quality of negative samples to balance scoring between positive and negative instances. Multi-modal data in KGs enhance traditional negative triple sampling (Bordes et al., 2013) by providing additional context for selecting higher-quality negative samples, thereby addressing a key performance bottleneck in KGC model training. Concretely, MMKRL (Lu et al., 2022b) introduces adversarial training to MKGC, adding perturbations to modal embeddings. This pioneers the use of adversarial methods for augmenting MKGC models. Following this, VBKGC (Zhang and Zhang, 2022) and MANS (Zhang et al., 2023f) develop fine-grained visual negative sampling to better align visual with structural embeddings for more nuanced comparison training. MMRNS (Xu et al., 2022c) introduces a relation-enhanced negative sampling method, utilizing a differentiable strategy to adaptively select high-quality negative samples.

FT-based Approaches leverage pre-trained Transformer models such as BERT (Devlin et al., 2019) and VisualBERT (Li et al., 2019), capitalizing on their profound multi-modal comprehension for MKGC. These methods transform MMKG triples into token sequences, feeding them into PLMs (Liang et al., 2022).

***(i)* Discriminative** strategies model KGC tasks as classification problems, with PLMs encoding textual information. KG-BERT (Yao et al., 2019), a forerunner in this field, fine-tunes BERT for triple classification, assessing triple plausibility based on the model's positive probability. Subsequent methods introduce additional tasks like relation classification and triple ranking (Kim et al., 2020; Wang et al., 2021a; Safavi et al., 2022), or explore prompt tuning in KGC (Lv et al., 2022; Chen et al., 2023a; Geng et al., 2023). FT-based MKGC methods emphasizes modal fusion over traditional KGC. Among them, MKGformer (Chen et al., 2022c) employs a hybrid Transformer for multi-level multi-modal fusion, treating MKGC as an MLM task and predicting masked entities by combining entity descriptions, relations, and images SGMPT (Liang et al., 2023a) extends MKGformer's capabilities by adding structural data integration through a graph structure encoder and a dual-strategy fusion module.

***(ii)* Generative** models frame KGC as a sequence-to-sequence task (Saxena et al., 2022; Xie et al., 2022; Chen et al., 2022a), employing PLMs for text generation. KGLLaMA (Yao et al., 2023b) and KoPA (Zhang et al., 2023i) explore the application of LLMs with instruction tuning for generative KGC, a relatively unexplored approach in MKGC, presenting a vast area for further exploration.

Discussion 4 *In MKGC, extracting modal information using pre-trained encoders like VGG or BERT is essential. Embedding-based approaches generally freeze these encoders during training and use the extracted data to initialize modal embeddings, while FT-based methods optimize them, aligning more closely with the model's inherent knowledge and memory. This leads to the underutilization of modal information in embedding-based methods, while FT-based methods struggle with complex KG structural information. Furthermore, the challenge of missing modal information in real-world KGs is* *significant. Initial solutions involved random initialization of missing modal embeddings, as seen in early works (Xie et al., 2017; Sergieh et al., 2018). Recently, MACO (Zhang et al., 2023h) introduce adversarial training to address this issue, but these methods remain basic, with a need for more innovative approaches.*

Resources & Benchmarks: ***(i)* Initial MKGC Datasets**: Early MKGC research primarily utilize established KG benchmarks such as WordNet (WN9-IMG (Xie et al., 2017), WN18-IMG (Wang et al., 2021b)), MovieLens100K (Pezeshkpour et al., 2018), YAGO-10 (Pezeshkpour et al., 2018), and FreeBase (FB) (Sergieh et al., 2018), extended with multi-modal information. For example, WN9-IMG incorporates images from ImageNet. ***(ii)* Systematic MKGC Datasets**: Liu et al. (2019b) transforms FB15K, DB15K, and YAGO15K into MMKGs by adding web-crawled images and numeric modal data. We benchmark those (M)KGC methods using this series of datasets as outlined in Table 7. Xu et al. (2022c) construct MKG-W and MKG-Y based on WikiData and YAGO, where the images are obtained through web search engines. ***(iii)* Multi-faceted MKGC Datasets**: Recent MMKGs include a broader range of modal information, represent the evolution towards more sophisticated datasets. For example, MMpedia (Wu et al., 2023b) is a scalable, high-quality MMKG developed using a novel pipeline based on DBpedia (Auer et al., 2007), designed to filter out non-visual entities and refine entity-related images through textual and type information. TIVA-KG (Wang et al., 2023h) spans text, image, video, and audio modalities, built upon ConceptNet (Speer et al., 2017). It introduces triplet grounding, aligning symbolic knowledge with tangible representations. In a similar vein, VTKG (Lee et al., 2023) attaches entities and triplets with images, supplemented by textual descriptions for each entity and relation.

A.7.1 Multi-modal Knowledge Graphs Reasoning

MKGC methods typically focus on single-hop reasoning in MMKGs, which may limit the exploitation of KGs for multi-hop knowledge inference (Das et al., 2018). Multi-modal knowledge graph reasoning (MKGR) aims to enable complex multi-hop reasoning on MMKGs, an area still in the early stages of research.

Table 7: Comparison of MKGC results on FB15K-237 and DB15K datasets (Liu et al., 2019b), with methods marked by † utilizing only text information for KGC with PLMs.

	Models	FB15K-237			DB15K		
		H@1	H@10	MRR	H@1	H@10	MRR
Embedding-based	IKRL (Xie et al., 2017)	.232	.493	.309	.111	.426	.222
	TBKGC (Sergieh et al., 2018)	.229	.494	.297	.108	.419	.208
	MKBE (Pezeshkpour et al., 2018)	.258	.532	.347	.235	.513	.332
	VBKGC (Zhang and Zhang, 2022)	.239	.478	.332	-	-	-
	MANS (Zhang et al., 2023f)	-	-	-	.204	.550	.332
	MoSE (Zhao et al., 2022c)	-	.565	.281	-	-	-
	MMRNS (Xu et al., 2022c)	-	-	-	.231	.510	.327
	HRGAT (Liang et al., 2023b)	.271	.542	.366	.597	.694	.630
	IMF (Li et al., 2023k)	.287	.593	.389	.427	.604	.485
	VISITA (Lee et al., 2023)	.287	.572	.381	-	-	-
FT-based	MTL-KGC† (Kim et al., 2020)	.172	.458	.267	-	-	-
	StAR† (Wang et al., 2021a)	.205	.482	.269	-	-	-
	SimKGC† (Wang et al., 2022c)	.249	.511	.336	.	-	.
	KGT5† (Saxena et al., 2022)	.210	.414	.276	-	-	-
	GenKGC† (Xie et al., 2022)	.192	.439	-	-	-	-
	KG-S2S† (Chen et al., 2022a)	.257	.498	.336	-	-	-
	CSProm-KG† (Chen et al., 2023a)	.269	.538	.358	-	-	-
	MKGformer (Chen et al., 2022c)	.256	.504	-	-	-	-
	SGMPT (Liang et al., 2023a)	.252	.510	-	-	-	-

Definition 5 ***MMKG Reasoning.*** *MKGR predicts a missing query element in one of three forms: $(h, r, ?)$, $(h, ?, t)$, or $(?, r, t)$, where "?" denotes the missing element. The objective is to infer this element through a multi-hop reasoning path in $\mathcal{T}_R$ of an* ***A-MMKG****, where the path length is shorter or equal to k hops, and k is an integer greater than or equal to 1.*

MMKGR (Zheng et al., 2023a) combines a gate-attention network with feature-aware reinforcement learning for multi-hop reasoning in MMKGs, guided by analogical examples. TMR (Zheng et al., 2023b) aggregates query-related topology features through an attentive mechanism to generate entity-independent features for effective MMKG reasoning under both inductive and transductive settings. MarT (Zhang et al., 2023b) introduces the concept of multi-modal analogical reasoning, akin to cross-modal link prediction but without explicitly defined relations. This task, framed as $(e_h, e_t) : (e_q, ?)$, leverages a background MMKG for missing element $(?)$ prediction. Its categorization under MKGR stems from its reliance on another triplet for tail (or head) entity prediction, differing from traditional MKGR in not requiring an explicit reasoning path. To facilitate this task, MarT presents a dedicated dataset (MARS) and an accompanying MMKG, MarKG. Additionally, they develop a model-agnostic baseline method inspired by structure mapping theory to address this unique reasoning challenge.

As this domain continues to evolve, it promises to become a pivotal direction in MMKG Inference, offering rich opportunities for groundbreaking discoveries and advancements.

A.8 MMKG-driven Tasks

Retrieval. As discussed in § 2, several MMKGs could naturally support retrieval related tasks: ImageGraph (Liu et al., 2017) connects a query to its top-K nearest neighbors, expanding via Bayes similarity-weighted edges up to a certain graph depth; IMGpedia (Ferrada et al., 2017), formatted in RDF, links visual descriptors and similarity relations with image metadata from DBpedia Commons, supporting SPARQL-based retrieval based on visual similarity, metadata, or DBpedia resources; VisualSem (Alberts et al., 2020) use a neural multi-modal retrieval model that processes both images and sentences to retrieve entities in the KG with pre-trained CLIP (Radford et al., 2021) as the encoder. Chen et al. (2021b) enhance MNER by searching the entire MMKG to acquire knowledge about poster images, using (mention, candidate entity) pairs from post text and MMKG for efficient image knowledge retrieval through iterative breadth-first traversal.

Cross-modal Retrieval. Zeng et al. (2023) provide a multi-modal knowledge hypergraph (MKHG) for linking diverse data in MMKGs and retrieval databases. a hyper-graph construction module with varied hyper-edges, multi-modal instance bagging for instance selection, and a diverse concept aggregator for sub-semantic adaptation, thus advancing representation learning in image retrieval. Huang et al. (2022) propose a unified continuous learning framework, iteratively updating the MMKG with MKGC as the target task and subsequently pre-training an MMKG-based VLM, using image-text matching as the core pre-training task without the need for paired image-text training data.

Reasoning & Generation. Zhao et al. (2021) introduce an Image Captioning method utilizing an MMKG that associates visual objects with named entities, leveraging external multi-modal knowledge from Wikipedia and Google Images for supplementary. The MMKG, after processing through a GAT (Velickovic et al., 2018), feeds its final layer output into a Transformer decoder, enhancing the precision of entity-aware caption generation. Jin and Chen (2023) involve the MMKG into multimodal summarization in a similar manner.

MMKG Pre-training. ***(ii)* Graph-level** Gong et al. (2023) aggregate various knowledge-view of the entities in MMKG (i.e., embeddings of neighbors connected by specific relations) to obtain their knowledge representation. These, combine with the entities' textual and visual embeddings, are integrated into CLIP's similarity computation process for multi-modal knowledge pre-training. Li et al. (2023j) introduce GraphAdapter for CLIP, a method that leverages dual-modality structure knowledge through a unique dual knowledge graph, comprising textual and visual knowledge subgraphs which represent semantics and their interrelations in both modalities. GraphAdapter enables textual features of prompts to utilize task-specific structural knowledge from both textual and visual domains, enhancing CLIP's classifier performance in downstream tasks.

AI for Science. AI for science refers to the application of AI techniques into scientific disciplines to drive discovery, innovation, and understanding. It employs AI to analyze, interpret, and predict complex scientific data, effectively supplementing traditional scientific methods with advanced computational tools. Within this domain, the concept of MMKGs is broadened beyond the conventional text and image modality to incorporate a diverse array of scientific data, including molecules, proteins, genes, drugs, and disease information (MacLean, 2021). This broader definition of "multi-modality" not only enriches the scope and depth of scientific research with varied data sources but also introduces new vitality and potential application value into the MMKG field.

In biology, MMKGs effectively integrate domain-specific data sources (Bonner et al., 2022) like Uniprot for proteins (Consortium, 2019), ChEMBL for small molecule-protein interactions (Gaulton et al., 2012), SIDER for side effects (Kuhn et al., 2016), and Signor for protein-protein interactions (Lo Surdo et al., 2023). These well-curated sources provide robust information to MMKGs. Additionally, data mined from extensive literature using NLP methods (Kilicoglu et al., 2012; Percha and Altman, 2018) further enrich MMKGs with diverse scientific insights. In those MMKGs, entities represent specific biological elements such as drugs or proteins, with relations depicting their experimentally verified interactions. These links, often augmented with additional attributes like molecular structures or external identifiers, can be directional to indicate causality, such as a drug causing a side effect (Ioannidis et al., 2020).

However, in the process of modeling complex biological systems, these MMKGs face challenges in MKGC due to data incompleteness, which hinders downstream applications. To address this, Xu et al. (2023d) create a co-attention-based multimodal embedding framework, merging molecular structures and textual data. It features a Triple Co-Attention (TCA) fusion module for unified modality representation and a relation-aware TCA for detailed entity-relation interactions, enhancing missing link inference. Moreover, biological MMKGs have also broadened their applications in **drug discovery**, extending beyond KGC to facilitate advanced tasks by leveraging rich graph knowledge. Lin et al. (2020) convert DrugBank data into an RDF graph using Bio2RDF, linking various biological entities and extracting triples for their KGNN framework. This framework predicts drug-drug interactions, adapting spatial-based GNN approaches to MMKGs by aggregating neighborhood information, which efficiently maps drugs and their potential interactions within the MMKG. Fang et al. (2022, 2023c) develop a chemical-oriented MMKG to summarize elemental knowledge and functional groups. They introduce an element-guided graph augmentation strategy for contrastive pre-training, exploring atomic associations at a microscopic level. Their approach, integrating functional prompts during fine-tuning, significantly improves molecular property prediction and yields interpretable results. Zhang et al. (2022a) construct a large-scale MMKG containing the Gene Ontology and related proteins. They implement a contrastive learning approach with knowledge-aware negative sampling to optimize MMKG and protein embeddings, enhancing protein interaction and function prediction. Cheng et al. (2023b) create an MMKG for protein science, integrating the Gene Ontology and Uniprot knowledge base. They develop a system for protein analysis, aiding predictions related to protein structure, function, and drug molecule binding, and supporting biological question answering. MMKGs thus serve not only as tools for direct query and pattern discovery but also as invaluable resources for augmenting and refining the performance of diverse computational tasks in drug discovery.

Remark 1 *Creating standardized benchmarks for biological MMKGs presents a challenge due to the varying sizes of these graphs and the diverse nature of the data they encompass. Despite these obstacles, several benchmarks have been developed to gauge progress in the field. OpenBioLink (Breit et al., 2020), for instance, is a benchmark specifically designed for large-scale biomedical link prediction. It provides a clear and transparent framework that facilitates the evaluation of new algorithmic approaches in this area. Additionally, PharmKG (Zheng et al., 2021c) has emerged as a dedicated benchmark specifically tailored for biomedical knowledge graph mining. Its introduction marks a significant step in advancing the field, providing researchers with specialized tools to evaluate and enhance data mining techniques in biomedical research. These benchmarks are crucial for the ongoing development and validation of computational methods, ensuring that innovations in MMKGs are both effective and relevant for practical applications in drug discovery. Zheng et al. (2021d) propose an MMKG attention embedding method for COVID-19 diagnosis, utilizing an image subset from public radiology reports and patient records, which contains three medical imaging modalities: X-ray, CT, and ultrasound. This offers a wider avenue for the future advancement of MMKG applications.*

Industry Application. Wang et al. (2023g) introduce FashionKLIP, a VLM enhanced by MMKG for **E-commerce**, incorporating FashionMMKG into a CLIP-style model for image-text retrieval. This approach uses contrastive learning for modal alignment and conceptual matching through visual prototypes from FashionMMKG for training.

MKGAT (Sun et al., 2020b) applies MMKGs to **movie and restaurant recommendation** systems, using a Collaborative MMKG (CMMKG) that merges user behavior with multi-modal item data. This model adopts entity-specific encoders and a GAT for entity representation, leveraging TransE for knowledge space learning. CKGC (Cao et al., 2022a) further categorizes traditional relations in MMKG into two types: descriptive attributes and structural connections, employing cross-modal contrastive learning for more effective node representation in recommendation.

B Future Directions

(i) As outlined in § 2, MMKG construction primarily involves two paradigms: annotating images

with KG symbols or grounding KG symbols to images. Recent developments, as highlighted in (Lee et al., 2023), start to explore a new path, **aligning locally extracted triples from multiple images with large-scale KGs**, which can be regarded as a mixture of MMKG and hyper-MMKG. The advantage of this hybrid approach is twofold: it not only extends the coverage of image quantity, as seen in the first paradigm, but also incorporates the extensive knowledge scale characteristic of the second. It promotes the generation of large-scale, triple-level multimodal information, posing both opportunities and challenges for future work in Multi-modal Entity Alignment and MMKG-driven applications like MLLM Pre-training and VQA.

(ii) Refining and aligning fine-grained knowledge within MMKGs is crucial. An ideal MMKG should be hierarchical, possessing deep levels with detailed and abstract multi-modal knowledge. Such a structure allows for the automatic decomposition of large-scale cross-modal data, enabling a single image to ground multiple concepts (Huang et al., 2023b). Moreover, segmentation represents an advanced requirement for grounding. With technologies like *Segment Anything* (Kirillov et al., 2023) already in place, such approaches can significantly reduce background noise impact in visual modalities. Thus, evolving towards **segmentation-level, hierarchical, and multi-grained** MMKGs marks a significant future direction.

(iii) In visual modalities, we hold that abstract concepts should correspond to abstract visual representations, while concrete concepts align with specific visuals. For example, general concepts like cats and dogs manifest in the brain as generic, averaged visual animal images, whereas specific qualifiers, such as "*Alaskan sled dogs*", provide clarity, similar to route-based image retrieval in MMKGs. Additionally, we also posit that every concept, visualizable or not, can be associated with certain modal representations. The abstract concept of "*mind*", for example, may evoke images of "*brains*" or "*people thinking*", still showing MMKGs' ability to represent NVCs. This perspective contrasts with previous views (Jiang et al., 2022; Peng et al., 2023). Interestingly, in human cognition, rarer concepts, such as "*unicorns*", are often more vividly depicted. If we know a *unicorns* only as a horned horse, this specific image is what we remember, rather than a horned seal or lion. This mirrors MMKG data structuring: concepts with fewer images are represented more distinctly, while those with more images are generalized and blurrier.

***(iv)* Efficiency in MMKG storage and utilization** remains a concern. Despite traditional KGs' lightweight nature and vast knowledge storage with minimal parameters, MMKGs demand more space, challenging efficient data storage and application across tasks. Enhancing efficiency might involve embedding multi-modal information into dense spaces as a temporary solution. Future research should strive to improve usage and storage efficiency without sacrificing MMKG's interpretability and structural integrity, a delicate balance that presents a continuing challenge.

***(v)* Quality control** in MMKGs introduces unique challenges with multi-modal (e.g., visual) content such as incorrect, missing, or outdated images. Limited fine-grained alignment between images and text in existing MMKGs and the noise from automated MMKG construction methods necessitate developing quality control techniques, possibly by assigning scores based on modal information quality. Given the dynamic nature of world knowledge, regularly updating MMKGs is essential. An important research direction lies in efficiently implementing **multi-modal knowledge conflict detection and updates**. The development of dynamic, temporal, and even spatiotemporal MMKGs (Liu et al., 2023d) is also crucial, enhancing their adaptability to diverse environments and user needs. Moreover, cross-lingual MMKGs can facilitate intercultural communication by enabling understanding and collaboration across languages and cultures, overcoming understanding barriers and supporting global cultural sharing.

MMKG for Tasks. Challenges in Scaling MMKG for Multi-modal Tasks: MMKG-driven tasks often emphasize retrieval-related activities, leveraging the natural database-like capabilities of MMKGs. However, the utilization of large-scale MMKGs in varied tasks, especially reasoning, is still nascent with limited exploratory studies. For example, Zha et al. (2023) enhance knowledge-based VQA by employing multi-modal concept descriptions and integrating MLLMs for refined answers. Nevertheless, these methods only use MMKGs as "*key:value*" based retrieval databases, not fully leveraging their multi-modal structured capabilities.

The constrained utilization of MMKGs in diverse tasks can be attributed to several factors. ***(i)* Non-Uniform Organization and Ontology** of MMKGs: Current MMKGs, lacking a standardized format, vary significantly in their focal points and the knowledge domains they cover for each downstream task. Predominantly, MMKGs cater to encyclopedic or trivia knowledge (Gong et al., 2023; Zhang et al., 2023a; Wu et al., 2023b; Zha et al., 2023), with commonsense and scientific related MMKGs (Wang et al., 2023h; Lee et al., 2023) being notably scarce. Moreover, the "non-visualizable" nature of some abstract knowledge components restricts their practical application (Jiang et al., 2022; Wu et al., 2023b). ***(ii)* Storage and Processing Overheads:** The substantial storage space requirements and extended processing times for large-scale MMKGs hinder their extensive adoption. Conversely, small-scale MMKGs frequently offer limited value for cross-task generalization. ***(iii)* Data Timeliness and Completeness Issues in MMKGs** heightens the risk of multi-modal hallucinations. ***(iv)* Comparative Advantages of LLMs and MLLMs:** LLMs and MLLMs excel in generalizability and AGI potential across various domains (Zhang et al., 2024), complementing the interpretability and editing flexibility of MMKGs. While MMKGs bring unique value, their development, maintenance, and application also involve certain costs. The evolving feedback from downstream tasks will continue to shape the industry's perspective on their respective roles and potentials.

Unlocking the Potential of Large-Scale MMKGs for Multi-Modal Tasks. *(i)* Integration with Non-text Modalities: Future downstream tasks driven by large-scale MMKGs can integrate methods from current KG-driven VQA methods, placing equal emphasis on non-textual modalities. This may further involve using modality projection or adapters for cross-modal alignment (Li et al., 2023j; Long et al., 2023), along with multi-modal GNN methods (Yoon et al., 2023) and modal feature decoupling techniques to enrich the granularity and hierarchy of multi-modal information (Chen et al., 2023g). ***(ii)* Rich Semantic MMKG Construction:** MMKG data can transcend traditional specialized or general formats. By developing task-specific pipelines, multi-modal datasets can be converted into MMKGs with enhanced semantics, using existing KGs as foundational references or bridges. This process can not only augments MLLM training with structured multi-modal input but also enriches the MMKG community with valuable, semantically rich datasets. ***(iii)* Reconstruction of Multi-Modal Tasks with LLM:** Combining LLM's text understanding and generation capabilities, multi-modal tasks can be restructured. Transforming KG-driven multi-modal tasks into in-MMKG-tasks, such as MKGC, MMEA, can enhance domain integration. There are already some attempts in this direction (Pahuja et al., 2024), which will be discussed in-depth later.

Large Language Models. The academic definition of LLMs, often associated with models possessing extensive parameters such as LLaMA-7B (Touvron et al., 2023), remains broad. These models' emergent abilities and Zero-shot Learning capabilities edge them closer to achieving AGI, underscoring their importance in NLP and multi-modal domains. The integration of multi-modal knowledge within LLMs, as seen in recent studies, prompts the semantic web community to delineate their distinct value amidst evolving (MM)KG-driven multi-modal methodologies.

***(i)* Fine-Tuning:** MMKGs provide a rich source of structured multi-modal data for Supervised Fine-Tuning (SFT) of MLLMs, especially in domain-specific applications (Zheng et al., 2024; Zhang et al., 2023g). Training techniques effective for MMKGs in VLMs can also be applied to MLLMs. The challenge of extracting sufficient visual knowledge, as identified by Chen et al. (2023b), alongside Zhou et al.'s (2023) finding that 43% of BLIP2 (Li et al., 2023e) errors on the A-OKVQA dataset (Schwenk et al., 2022) could be addressed with proper knowledge integration, emphasizes the need for embedding explicit and especially long-tail knowledge into MLLMs (Zhang et al., 2023c). This process within MMKGs can be realized along two distinct pathways: one involves active KG routing exploration for constructing specific instructions (Wan et al., 2023), and the other leverages self-instructing techniques to autonomously evolve and generate multi-grained, multi-modal instructional data (Wang et al., 2023i; Xu et al., 2023b; Du et al., 2023a; Yona et al., 2024). Besides, the structured multi-modal relational data inherent in MMKGs provides an essential foundation for investigating the visual extrapolation abilities of purely visual LLMs, or Large Vision Models (LVMs) (Bai et al., 2023), as well as MLLMs (Sun et al., 2023d; Wei et al., 2023a). Furthermore, MMKG data can be utilized to further explore the concept of multi-modal reversal curse (Lv et al., 2023), where the ordering of knowledge entities in training data influences model comprehension, potentially limiting the model's understanding.

***(ii)* Hallucination:** As LLMs rapidly advance, the risk of generating seemingly authentic but factually inaccurate web content is increasing. This phenomenon, known as *hallucination* (Zhang et al., 2023j; Rawte et al., 2023; Agrawal et al., 2023), often arises from outdated or incorrect training encountered during the model training process, or from the frequent co-occurrence bindings of objects, affecting both LLMs and MLLMs (Huang et al., 2023a; Tong et al., 2024; Liu et al., 2024a). To combat this, LAMM (Yin et al., 2023) incorporates 42K KG facts from Wikipedia and leveraged the Bamboo dataset (Zhang et al., 2022c) to refine commonsense knowledge in Q&A, underscoring the role of quality (MM)KGs in mitigating LLM hallucinations (Agrawal et al., 2023; Xu et al., 2023f). Developing robust hallucination detectors (Chen et al., 2023c; Mishra et al., 2024) is crucial for identifying and curbing errors in LLM outputs. Future efforts could focus on pairing MMKGs with detection methods to improve multi-modal task precision and leveraging (MM)KGs for knowledge-aware statement rewriting to diminish factual hallucinations in LLM reasoning (Guan et al., 2023; Wang et al., 2023b).

***(iii)* Agent:** Multi-agent Collaboration (Xu et al., 2023e; Xiao et al., 2023; Lu et al., 2024), simulating human cognitive processes, can dissect VQA reasoning paths and engage multiple (M)LLMs in collective problem-solving (Wang et al., 2023j; Qiao et al., 2024). In this framework, KGs can initialize agent personalities (Mao et al., 2023; Tu et al., 2023), providing a structured basis for intuitively designing character brains, enriching the interaction between agents and enhancing their collective reasoning capabilities.

Chain-of-thought (CoT) reasoning (Wei et al., 2022) significantly improves LLMs' complex reasoning abilities by incorporating intermediate reasoning steps. This progress has catalyzed the emergence of various KG-focused applications (Park et al., 2023; Sun et al., 2023b). For example, Sun et al. (2023b) demonstrate how LLMs can be used to interactively navigate KGs to extract knowledge for reasoning. Their Think-on-Graph (ToG) approach utilizes beam search to identify effective reasoning paths within KGs. Merging these innovations with MMKGs promises to expand the scope of tasks, especially in improving the ability of models to interpret and interact with diverse data types, such as images and text (Mondal et al., 2024). This integration moves us closer to achieving human-like multi-modal proficiency and paves the way for advanced machine intelligence.

***(iv)* RAG:** Retrieval Augmented Generation (RAG) (Ovadia et al., 2023) systems enhance (M)LLMs by incorporating long-tail knowledge beyond their parameter limits. However, excessive document retrieval can lead to contextually inappropriate answers (Barnett et al., 2024), increasing hallucination risks unless carefully designed prompts are used (Wang et al., 2023k). The high information density and structured organization in KGs can mitigate this issue. Moreover, MMKGs can further aid multi-modal RAG by using various modalities as anchors (Song et al., 2023a), offering more relevant and explanatorily powerful results than vector-based searches (Wu and Xie, 2023; Yu et al., 2023).

***(v)* MMKG Refinement:** LLMs offer the capability to augment MMKGs through their advanced text comprehension and generation skills. Recent work, such as (Yao et al., 2023b; Zhang et al., 2023i), explores LLM-based KGC. Specifically, KoPA (Zhang et al., 2023i) integrates KG structural knowledge into LLMs to enable structure-aware reasoning. Moreover, with the continuous growth and evolution of online data, LLMs can support the continuous learning and self-updating of MMKGs, serving as active annotators (Zhang et al., 2023d).

***(vi)* MMKG MoE:** The Mixed of Expert (MoE) architecture shows outstanding performance in LLM applications. Initially, it engages input samples through a GateNet or router for multi-class categorization, determining the allocation of tokens to appropriate experts. This critical process, known as *experts selection*, is central to MoE's concept, often characterized as sparse activation in academia (Ismail et al., 2023; Dou et al., 2023; Team, 2023; Dai et al., 2024; Lin et al., 2024). These experts then process the inputs to formulate final predictions. Regarding domain-specific MMKGs in fields like biology, e-commerce, and world geography, an innovative direction involves creating an extensive MMKG library (or repository). This library would house varied MMKGs, each tailored to specific domains, allowing downstream tasks to adaptively select relevant MMKG information in a manner akin to MoE's. Exploring this conceptual approach could not only catalyze developments in MMKG-level retrieval and re-ranking but also foster the seamless integration of MMKGs into model parameters, merging their utility with the dynamic allocation efficiency of MoE architecture.

AI for Science. Despite the vast potential of biological MMKGs in drug discovery, several challenges exist. One of the primary hurdles is the issue of data heterogeneity and quality, which demands meticulous integration and standardization of data from diverse sources. Another major challenge lies in the choice of knowledge representation within these MMKGs. The ideal representation would capture the intricate details of drug discovery objects and relationships, such as the various protein isoforms produced from a single gene and their complex interactions within cellular environments. However, achieving this level of detail is often hindered by practical constraints like cost and technology limitations. Furthermore, specific data sources may impose additional limitations based on their existing information structures. As such, the chosen knowledge representation in MMKGs often has to strike a balance between desired granularity and practical feasibility, reflecting both the current state of scientific knowledge and the inherent limitations of data sources. This balancing act poses a significant challenge and indicates the need for ongoing efforts to refine and expand the scope and depth of MMKGs in the field of drug discovery.