Knowledge Graphs for Multi-modal Learning: Survey and Perspective

Anonymous ACL submission

Abstract

Integrated with multi-modal learning, knowledge graphs (KGs) as structured knowledge repositories, enhance AI's capability to process and understand complex, real-world data. This paper provides a comprehensive survey of cutting-edge research on KG-aware multi-modal learning, providing task definitions, evaluation benchmarks, and detailed insights into key breakthroughs. Furthermore, we also discuss current challenges, highlighting emerging trends and future research directions.

1 Introduction

The brain's accumulation of memories over time is crucial for societal adaptation and survival, enabling meaningful actions and interactions. These memories can be categorized into two types. *Conditioned Reflexes:* Intuitive memory developed through repeated practice, enhancing analogical reasoning. Combined with sensory inputs like visual, auditory, and tactile data, it enables efficient performance of basic tasks, similar to many traditional multi-modal tasks. *Torso-to-tail Knowledge:* Less frequently encountered, requiring active memorization or deep contemplation. This knowledge highlights the importance of Knowledge Graphs (KGs) in capturing and structuring long-tail knowledge. Our study focuses on leveraging symbolic, structured knowledge within KGs to enhance Multi-Modal Learning (MML). Given their vital role in organizing long-tail knowledge and proven effectiveness in AI systems, integrating KGs with MML offers a promising approach to addressing the challenges inherent in multi-modal data integration.

As illustrated in Fig 1, individuals continuously process multi-modal information from the environment while absorbing and utilizing external knowledge. Despite extensive research within NLP communities, a systematic review of these approaches remains absent. Our survey fill this gap by synthesizing existing KG4MML methods, detailing key resources and breakthroughs. We balance these details to address content overlaps and focus on core challenges, ultimately emphasizing KG's pivotal role in shaping the past, present, and future developments of multi-modal learning.

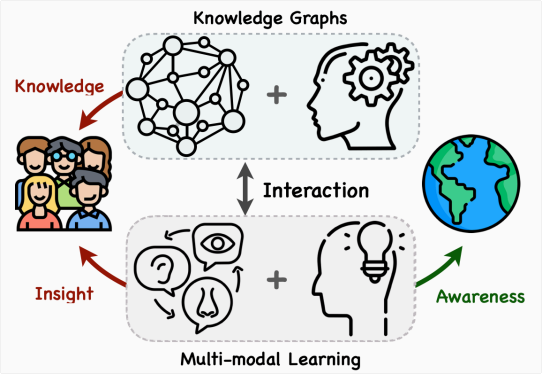


Figure 1: Knowledge Graphs for Multi-modal Learning.

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), where h is the head entity, r is the relation, and t is the tail entity, or attribute triples (e, a, v), where e is an entity, a is an attribute, and v is the attribute's value. Attribute values can be literals such as strings or dates and may include metadata like labels and textual definitions.

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

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

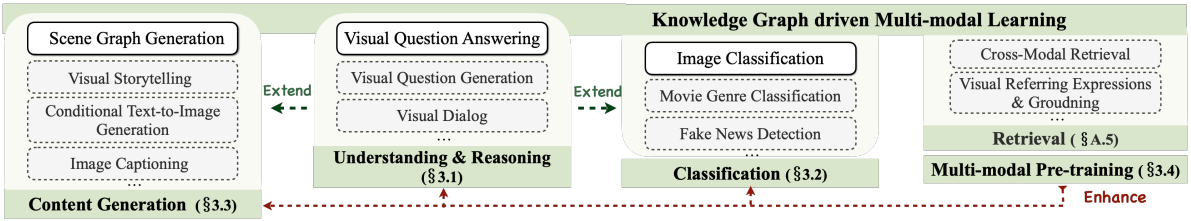


Figure 2: Comprehensive Overview of KG-driven Multi-modal Learning. Due to space constraints and task overlaps, we focus on the most representative sub-tasks in each category (Generation, Understanding & Reasoning, Classification) to maximize relevant content coverage. Additional content is analyzed in the Appendix[1].

be adapted to them. Thus, the input domain is $\mathcal{X} = \mathcal{X}^{\mathbb{I}} \times \mathcal{X}^{\mathbb{V}}$, with inputs $\hat{x} = (x^{\mathbb{I}}, x^{\mathbb{V}})$, where $x^{\mathbb{I}}$ and $x^{\mathbb{V}}$ are language and visual data, respectively.

KGs Acquisition for Multi-modal Learning. *1) Task-agnostic Sub-KG Extraction:* Practical applications often require localized knowledge to address specific tasks. Sub-KG extraction isolates minimal knowledge units or triplets from a large KG like WordNet (Miller, 1995) directly, reducing noise from irrelevant information using retrieval, routing, or semantic parsing algorithms.

2) Task-Oriented KG Construction: Sometimes, constructing task-specific KGs from scratch is necessary to meet unique requirements, either from datasets or by combining multiple KGs: ***(i)*** Static Domain KGs Construction: Creating stable, domain-specific KGs with predefined entities and relations to encapsulate crucial background knowledge, especially when general KGs are inadequate for a specific task. E.g., Zero-shot Image Classification tasks necessitate building KGs that capture visual attributes or taxonomy associations (Geng et al., 2021a). These KGs are designed to cover all relevant classification knowledge with textual data like class labels utilized to delineate class relationships, aiding in the formation of KG edges. ***(ii)*** Dynamic Temporary KGs Construction: Building dynamic, temporary KGs during task execution and leveraging KG reasoning algorithms for task support. E.g., establishing co-occurrence relations between classes (e.g., food ingredients) by analyzing their frequency in training datasets.

3 KG-driven Multi-modal Learning

3.1 Multi-modal Understanding & Reasoning

Visual Question Answering (VQA) (Antol et al., 2015) is a fundamental task in multi-modal learning, frequently used to evaluate multi-modal foundation models (Alayrac et al., 2022). Knowledge-based VQA (Wu et al., 2016) incorporates an external Knowledge Base (KB) for complex question analysis and deeper reasoning assistance (Wang et al., 2018a). When using KGs as the KB, given an image-question pair (I, Q), the goal of KG-based VQA is to derive an answer y via:

$$p(y|Q, I, \mathcal{G}, \mathbf{\Theta}) = \underbrace{p(\mathcal{G}_{ret}|Q, I, \mathcal{G}; \mathbf{\Phi})}_{\text{Retriever (if have)}} \cdot \underbrace{p(A|Q, I, \mathcal{G}_{ret}; \mathbf{\Theta})}_{\text{Reader}},$$

where $\mathcal{G}$ is the background KG, $\mathcal{G}_{ret}$ is the retrieved sub-KG, while $\mathbf{\Phi}$ refers to the model parameters used in the knowledge retrieval step.

Current KG-aware VQA research typically has four key stages for incorporating knowledge:

- **Knowledge Retrieval** focuses on extracting pertinent knowledge from various sources, including KGs and document collections like Wikipedia (Denoyer and Gallinari, 2006). The methods have evolved from early matching-based and dense embedding similarity approaches to learnable retrieval and Pre-trained Language Model (PLM) generation techniques, broadening the scope and efficiency of knowledge integration. Fig. 4 (a) illustrates the basic form of KG retrieval in VQA.

(i) **Matching-based Retrieval** generally employs entity-level methods to identify key concepts within images and questions, linking these to relevant data to external large-scale KGs.

The extraction process from images involve identifying spatial positions (Zhu et al., 2020), visual object sizes and names (Narasimhan et al., 2018), high-level attributes like scene names, object parts, and human activities (Khademi et al., 2023). Additionally, image captions and OCR text strings are generated for supplement (Lin and Byrne, 2022). Questions and captions can be parsed for syntax analysis using various NLP tools like Dependency Parsers and Named Entity Recognizers (Wu and Mooney, 2022). Additional techniques include regular expressions (Wang et al., 2017), SpanSelector or query template selector (Wang et al., 2018a). During this stage, unimportant visual objects not present in the question or caption might be filtered out (Gardner et al., 2018). To associates these concepts with relevant entries in KGs, methods like greedy longest-string matching (Su et al., 2018),

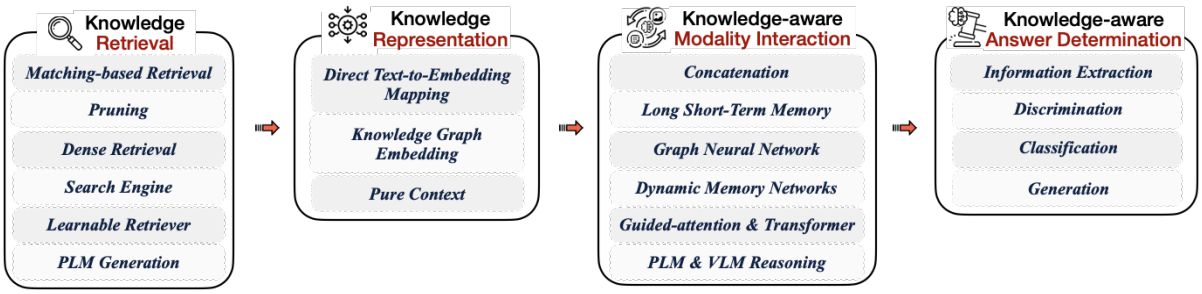


Figure 3: Current KG-aware Understanding & Reasoning research pipeline, which typically involves four key stages for incorporating knowledge. Note that studies often employ one or more of these stages in model design.

template matching (Wang et al., 2018a), and Multimodal Entity Linking (Jain et al., 2021) are used. Then, fact triples can be collected by involving the first-order sub-KG from these identified concept nodes, which can sometimes extend to three-hop connections in character KGs (Shah et al., 2019). Alternatively, brief knowledge paths can be identified among the entities from I and Q via sub-KG construction (Wang et al., 2019). Generating RDF queries (e.g., SPARQL) often involves filling predefined templates with parsed question data, making it suitable for datasets with consistent question patterns (Wang et al., 2017). Term-based retrievers (e.g., TF-IDF and BM25) are another effective option, with their scoring reflecting the direct correlation between Q and fact triples (Luo et al., 2021).

***(ii)* Pruning.** The pruning stage refines the coarse-grained sub-KG obtained from initial retrieval. This involves re-ranking candidate facts and may include assigning weights to nodes based on corresponding visual object sizes (Wang et al., 2019), ensuring each knowledge triple contains key elements from Q or auto-generated captions (Su et al., 2018; Wu and Mooney, 2022), or aligns with the relation type implied in Q (Yin et al., 2023a). Additionally, a learnable score function can be used to assess the compatibility between a fact and the Q-I representation (Ravi et al., 2023).

***(iii)* Dense Retrieval** methods typically retrieve the most relevant top-k facts for a given Q-I pair. This technique utilizes embedding similarities to match questions and visual concepts with preflattened concise fact sentences (Narasimhan et al., 2018; Ziaeefard and Lécué, 2020), simplifying the retrieval process without complex rules. Sometimes, dense retrieval can also serve as a mechanism for KG pruning, selectively excluding information that is likely irrelevant. Retrieval efficiency is frequently enhanced by employing open-source indexing engines like FAISS (Johnson et al., 2021), which facilitates the organization and indexing of large-scale dense embeddings.

***(iv)* Search Engine** serves as a valuable tool in VQA for accessing open knowledge and can complement other knowledge sources in ensemble methods. For instance, Marino et al. (2019) gather Wikipedia articles for each Q-I pair, selecting sentences that closely match the query by keyword frequency and then predicting answer presence and positioning within these articles. Given that this is not a primary focus for KG-based VQA, a more detailed discussion is provided in the appendix.

***(v)* Learnable Retrievers** adapt to specific contexts, providing biased recall that highlights interactions between Q-I and knowledge components. These models require rigorous training, using either labeled data or autonomously generated labels. For example, UnifER (Guo et al., 2022b) compares reader loss from Q-I inputs to loss with additional retrieved knowledge, defining the difference as the *loss gap*. A negative *loss gap* indicates counterproductive knowledge, and the model uses this metric to iteratively refine both the retriever and the reader. **Cold Start** issues with asymmetric or randomly initialized retrievers can lead to irrelevant items and inadequate feedback. Hu et al. (2023) address this by creating an initial dataset with pseudo ground-truth knowledge from a large-scale image-caption dataset for pre-training. Chen et al. (2022c) use a BM25-based KG retriever initially to distill preference knowledge to the differentiable retriever.

***(vi)* PLM Generation.** Recent research shows that PLMs can function as KBs when appropriately prompted (Petroni et al., 2019). Specifically, PROOFREAD (Zhou et al., 2023) uses ChatGPT to generate relevant Q and knowledge entries for each Q-I pair, storing them in a *Demo Bank* for demonstration reuse while ensuring case diversity; Additionally, several studies directly harness the knowledge embedded in PLMs for reasoning, skipping a separate knowledge retrieval step (Yang et al., 2022). E.g., CodeVQA (Subramanian et al., 2023)

involves a knowledge query module that utilizes a PLM to answer questions based on world knowledge embedded within its parameters.

• **Knowledge Representation** involves selecting the appropriate format for symbolic KGs to integrate with multi-modal models.

***(i)* Direct Text-to-Embedding Mapping.** Some research treats e and r in KGs as words, embedding them into continuous vectors using methods like Glove. This enables the compression of knowledge components (e.g., triples) into fixed-size vectors using RNNs (Wang et al., 2019), (V)PLMs (Chevalier et al., 2023), or mean pooling (Chen et al., 2021d). Strategies like stop-word removal can further refine these embeddings by reducing noise from irrelevant words (Chen et al., 2022c). Additionally, some works convert fact collections into natural language sentences (Hu et al., 2023), allowing PLMs to encode them directly into fixed-length vectors.

***(ii)* Knowledge Graph Embedding** (KGE) methods embeds facts and reveals semantic relationships among triples in an abstract space, facilitating initial fact embeddings (Zheng et al., 2021b; Han et al., 2023). During self-supervised training, signals from neighboring entities are embedded into each central entity's unique representation. This approach facilitates the identification and integration of key entities into downstream tasks, effectively simulating the retrieval of specific sub-KGs without explicit retrieval steps.

***(iii)* Pure Context.** In many cases, KG triples are maintained in their original textual format for direct participation in multi-modal reasoning. This includes using sub-KGs for RDF query-based answer retrieval (Wang et al., 2017) and serializing triples for joint reasoning with (V)PLMs (Gao et al., 2022; Dong et al., 2024). To handle lengthy input sequences predominantly composed of facts, which might shift the model's focus away from other crucial cues, Ravi et al. (2023) summarizes information from each inference sentence into a single token representation using SBERT (Reimers and Gurevych, 2019); Wang et al. (2023g) select only factual summaries with higher contribution scores than the original Q as caption supplements.

• **Knowledge-aware Modality Interaction** is the core of KG-based multi-modal reasoning. To some extend, it mirrors human knowledge application in understanding the world (Fig. 4).

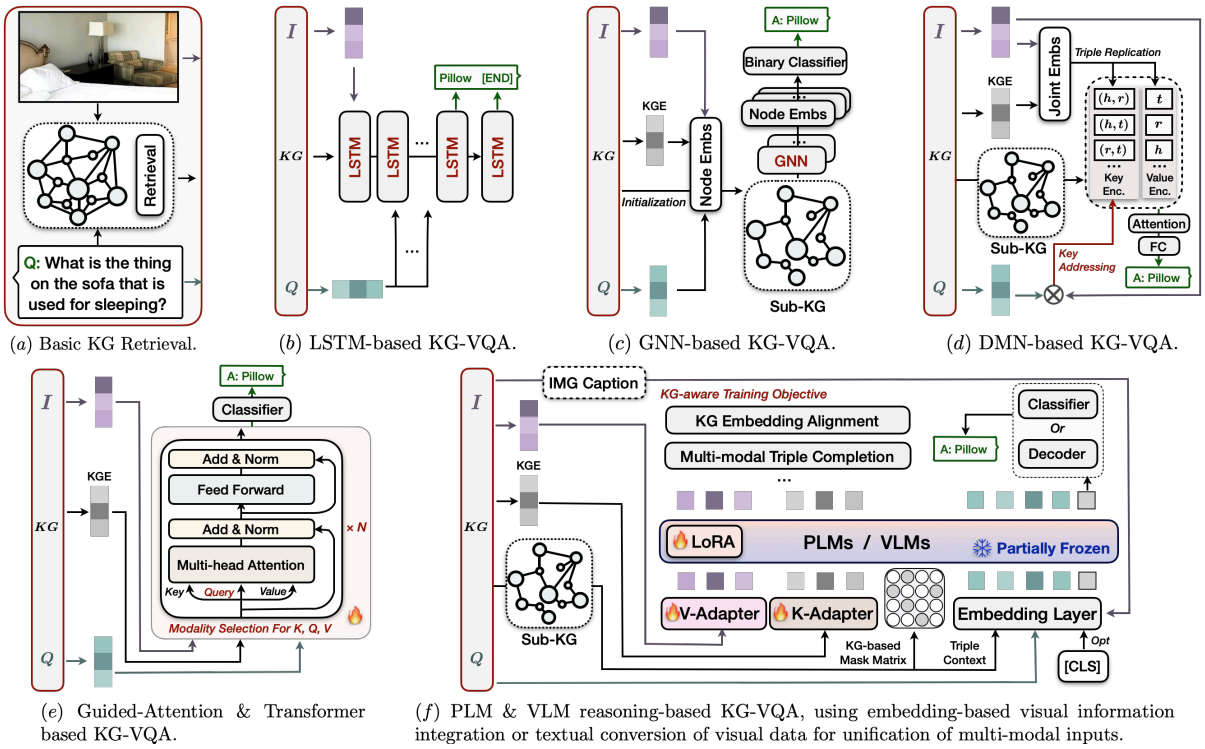
***(i)* Concatenation** of multi-modal vectors provides a straightforward and effective approach for modality fusion (Ramnath and and, 2020), combining different modality features into a single representation. This unified feature is typically refined with a MLP to enhance modality interaction and integration.

***(ii)* LSTM** networks typically employ an LSTM encoder to process semantic inputs from I and Q, and an LSTM decoder for generating answers, initializing the hidden state with embeddings from attributes, captions, or external knowledge. As shown in Fig. 4 (b), Q is tokenized and fed into the system sequentially (Wu et al., 2016).

***(iii)* GNNs** enhance concept connections in VQA by integrating representations from I, Q, and entities into cohesive networks, with each node (entity e) represented by a multi-modal concatenated embedding (Narasimhan et al., 2018). GNNs iteratively process these e embeddings, and the final learned representations are fed into an MLP, which assigns a binary label to each e indicating its relevance as an answer, as shown in Fig. 4 (c). Mucko (Zhu et al., 2020) independently processing distinct modality KGs, separately analyzing the visual scene KG, the semantic KG from image captions, and the common sense KG, which supports precise answer determination through Q-guided attention and cross-KG GNNs.

***(iv)* Dynamic Memory Networks** (DMNs) (Kumar et al., 2016) filter critical information from localized small-scale knowledge triple embeddings (Fig. 4 (d)) to enable interactions across multiple data channels (Yin et al., 2023a). This process, akin to building a cache and performing secondary retrieval, is typically achieved through an attention-based mechanism. In particular, VKMN (Su et al., 2018) deconstructs each knowledge triple into three Key-Value pairs, e.g., (h, r) as the key and t as the value, improving reasoning performance by reducing interference from using only head and tail entities as keys for memory grounding.

***(v)* Guided-Attention & Transformer.** The Transformer architecture, featuring multi-head attention, layered stacking, and residual connections, is widely used in multi-modal fusion (Vaswani et al., 2017). It enables knowledge embeddings to integrate seamlessly with other modalities. The guided-attention mechanism (Heo et al., 2022) enhances this by using distinct feature sets to direct attention unidirectionally, unlike self-attention's symmetric interactions. This is intuitive, reflecting the unequal roles of different modalities and knowledge in fusion. Examples include knowledge-



(a) Basic KG Retrieval. (b) LSTM-based KG-VQA. (c) GNN-based KG-VQA. (d) DMN-based KG-VQA.

(e) Guided-Attention & Transformer based KG-VQA.

(f) PLM & VLM reasoning-based KG-VQA, using embedding-based visual information integration or textual conversion of visual data for unification of multi-modal inputs.

Figure 4: Current knowledge-aware modality interaction paradigms utilizing KG as the knowledge repositories.

guided visual/textual embeddings and Q-guided visual/knowledge embeddings (Fig. 4 (e)).

***(vi)* PLM & VLM Reasoning** allow researchers to focus on data organization and training objective design without significantly altering the backbone structure, effectively utilizing the inherent parameterized knowledge in original models (Fig. 4 (f)):

(a) **Embedding-Based Visual Information Integration** involves converting visual data into embeddings compatible with existing (V)PLMs, like compressing patch or object features into fixed-length embeddings (Jaegle et al., 2021) or using adapters/projection heads for cross-modal alignment (Yin et al., 2023b). Specifically, RVL (Shevchenko et al., 2021) and KVQAmeta (García-Olano et al., 2022) inject the knowledge into VLMs via aligning the KG embeddings of e with corresponding textual phrase representations derived from the PLM's embedding layers. MuKEA (Ding et al., 2022) uses VLM's visual output embeddings as the head, question as the relation, and the answer as the tail to design a multi-modal triple completion training objective, leveraging implicit knowledge for reasoning.

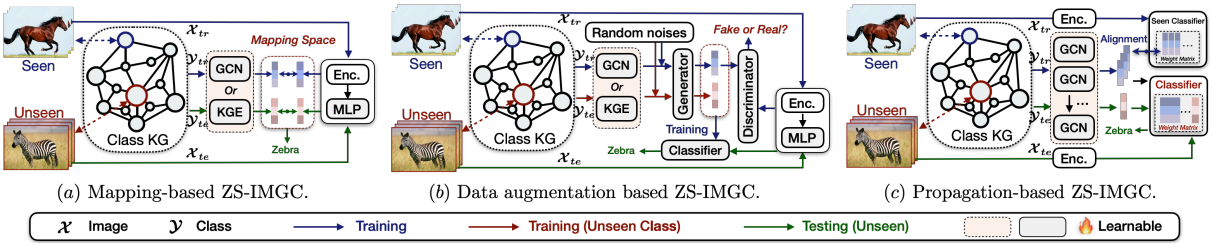
(b) **Textual Conversion of Visual Data** involves converting all visual information into a textual format, like captions, allowing PLM reasoning on a uniform textual data collection that includes background knowledge, Q, and I (Hu et al., 2022).

- **Knowledge-aware Answer Determination.**

***(i)* Information Extraction** methods typically focus on locating specific entities within the retrieved knowledge or existing I-Q pair, emphasizing content grounding. Among them, query-based methods (Wang et al., 2017) obtain the final answer through sub-KG inference, yielding benefits such as improved matching accuracy, relevance, and interpretability. Their effectiveness, however, depends on the model's capability to parse queries and the KG's completeness. Challenges arise with non-unique or difficult-to-find answers. To further rank the potential answers, heuristic rules can be employed, like matching score calculation (Wang et al., 2018a; Narasimhan and Schwing, 2018) and answer frequency assessment (Wang et al., 2018a).

***(ii)* Discrimination.** Particularly suited in multi-choice VQA tasks, these methods use a discriminator for final selection among candidate answers. They are effective in narrowing down potential answers within a certain range (or in a sub-KG), often using GNN-alike backbones (Hussain et al., 2022) (Fig. 4 (c)). Furthermore, discriminators can be either MLP-based (Liu et al., 2022) or rule-based (Narasimhan and Schwing, 2018), with a notable limitation being time consumption, especially with extensive answer vocabularies.

***(iii)* Classification.** In many VQA datasets, the range of possible answers is pre-determined, typically constrained by their frequency range or a minimum occurrence threshold. Consequently, many



(a) Mapping-based ZS-IMGC. (b) Data augmentation based ZS-IMGC. (c) Propagation-based ZS-IMGC.

Figure 5: A comparison of previous paradigms for KG-based Zero-shot Image Classification (ZS-IMGC) methods.

studies reformulate VQA as a classification problem (Gardères et al., 2020; Song et al., 2023b), with the output dimension corresponding to the number of answer candidates. For (V)PLM-based methods, a classification (or projection) head is typically appended to the output `[CLS]` embedding, as shown in Fig. 4 (f) (He and Wang, 2023).

***(iv)* Generation.** With the expansion of parameters and pre-training data in LMs (Schaeffer et al., 2023), generative models' accuracy has significantly improved, effectively mitigating the disadvantages posed by the exact match criteria that constrained early LSTM-based methods. As a result, generative (V)PLM-based methods are now increasingly supplanting traditional classification-based approaches (Ghosal et al., 2023) (Tab. 1).

3.2 Multi-modal Classification

This section explores multi-modal classification tasks, particularly focusing on Zero-Shot Image Classification (**ZS-IMGC**), which classifies images from novel, unseen classes without prior specific training, as denoted by $\mathcal{Y}_{tr} \cap \mathcal{Y}_{te} = \emptyset$. In contrast, traditional image classification (x as an image and y as its class) assumes a closed world where $\mathcal{Y}_{tr} = \mathcal{Y}_{te} = \mathcal{Y}$, demanding extensive labeled images for both training and testing within known classes.

Early ZS-IMGC works use textual class characteristics (Zhu et al., 2018) and define shared descriptive attributes (Xian et al., 2018) to model inter-class relationships (see Fig. 10, left). Recently, KGs have become integral to ZS-IMGC by unifying various forms of above knowledge into a single graph, $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, where $\mathcal{Y} \subset \mathcal{E}$. This integration not only encapsulates diverse and explicit class semantics but also enhances compatibility and interpretability (see Fig. 10, right). Specifically, studies like (Kampffmeyer et al., 2019) integrate hierarchical relationships from WordNet, while Roy et al. (2022) explore class knowledge from commonsense KGs (e.g., ConceptNet). Pahuja et al. (2024) enhance species classification by structuring it as a link prediction task within a Multi-Modal Knowledge Graph (MMKG), leveraging visual cues and GPS coordinates to efficiently identify unseen classes, such as deducing that an image of a feline captured in Africa is likely a lion. Furthermore, ontologies can be utilized to define complex class relationships (Chen et al., 2020a) (e.g., *disjointness*), providing a channel to explicitly introduce rules for refinement.

Existing KG-driven ZS-IMGC approaches, which guide feature transfer from seen to unseen classes, can be categorized into three types:

Mapping-based methods develop mapping functions that align image inputs with KG-based class semantics into a shared vector space, simplifying the identification of the nearest class to a test image based on similarity metrics, as shown in Fig. 5 (a). For instance, HierSE (Li et al., 2015) employs a linear projection to map image features into a class embedding space derived from word embeddings of the class and its superclasses, using cosine similarity for comparison. Similarly, Chen et al. (2020a) use an OWL-based ontology to encode animal classes, and Akata et al. (2013) represent classes using multi-hot vectors of ancestors, reflecting class hierarchies in KGs.

Data Augmentation methods alleviate the sample shortage in ZS-IMGC by synthesizing images or features for unseen classes, transforming it into a supervised learning problem and reducing bias. These methods primarily use generative models like GANs and VAEs, leveraging feature-related KGs to simulate characteristics of unseen images. For example, OntoZSL (Geng et al., 2021a) synthesizes image features by blending class embeddings from an attribute-and-species KG with random noise vectors. This process, supervised by real features from annotated images, employs an adversarial discriminator to differentiate between real and generated features as shown in Fig. 5 (b).

Propagation-based methods leverage KG-structured inter-class relationships for knowledge transfer, using GNNs to propagate features from seen to unseen class nodes (Wang et al., 2018b; Geng et al., 2021b). Concretely, GNN models train to produce class-specific parameter vectors as

classifiers. These classifiers for unseen classes are estimated by aggregating those from neighboring seen classes within the graph (see Fig. 5 (c)).

3.3 Multi-modal Generation

Given visual images ($x^{\mathbb{V}}$) or textual descriptions ($x^{\mathbb{L}}$), the objective of KG-aware generation tasks is to generate, in a cross-modal manner, either a textual target $y^{\mathbb{L}}$ (e.g., caption and scene graph), a visual target $y^{\mathbb{V}}$ (e.g., image), leveraging the background KG $\mathcal{G}$ for foundational support[2].

Scene Graphs (SGs) are vital for scene understanding, cataloging objects and their interrelationships within a scene. These instances, ranging from people to places and objects, are described through attributes like shape, color, and pose (Chang et al., 2023). The relationships between these instances, often action-based or spatial, are expressed as *(subject, predicate, object)* triplets, paralleling the (h, r, t) and (e, a, v) triplets in KGs. Scene Graph Generation (SGG) serves as an intermediary task, unlike other multi-modal tasks with specific end goals, providing enhanced understanding and reasoning to support downstream tasks (Fig. 6).

Several works adopt the **KG representation learning** techniques into SGG scenario. For example, Yu et al. (2022) improve zero-shot performance in SGG by constructing a KG from training set SG triples, distinguishing existing (non-zero-shot) and missing (zero-shot) edges. They train a KG Embedding model to complete the graph and fills these missing edges, thereby integrating zero-shot triples similarly to their non-zero-shot counterparts. Others employ KGs for **triple prediction** to generate rich and expressive SGs. Specifically, Gu et al. (2019) utilize a knowledge-based module to identify relevant ConceptNet entities and retrieve commonsense facts, each assigned a weight indicating its real-world prevalence to filter candidate triples. Khan et al. (2022b) enrich SGs using CSKG (Ilievski et al., 2021), a substantial commonsense KG repository as shown in Fig. 6. This upgrades SGG with additional information on objects' spatial proximity and potential interactions derived from external knowledge, improving higher-level reasoning and mitigating some missed or incorrect predictions made during SGG.

[2]We defer detailed discussion to Appendix A.5, noting significant overlaps with reasoning tasks like Image Captioning and Visual Storytelling, similar to VQA as discussed in § 3.1, and because KG-aware cross-modal image generation, where only limited work exists, is still in its early stages.

[3]Off-scene entities refer to those not part of the VG (Krishna et al., 2017) classes, as opposed to on-scene entities.

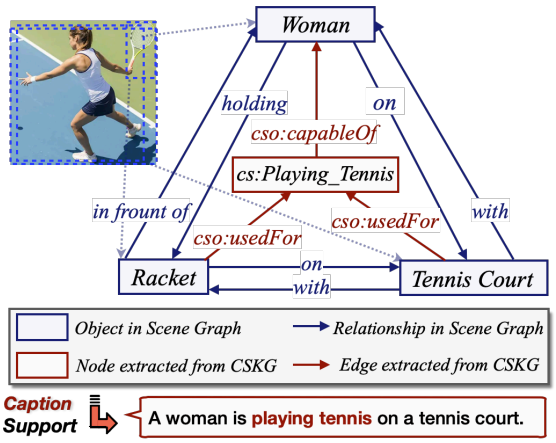
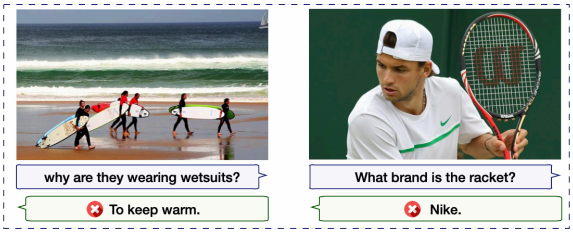


Figure 6: Scene Graph (SG) of an image equiped with CSKG (Ilievski et al., 2021). The SG (in blue) outlines objects and their relationships in the scene. Additional background knowledge from CSKG triples (in red) such as *(Woman, capableOf, Playing_Tennis)*, enriches the SG with off-scene[3] knowledge (Khan et al., 2022b). This facilitates higher-level inferences, enabling more accurate caption deductions, e.g., "*A woman is playing tennis on a tennis count*" (Hou et al., 2020).

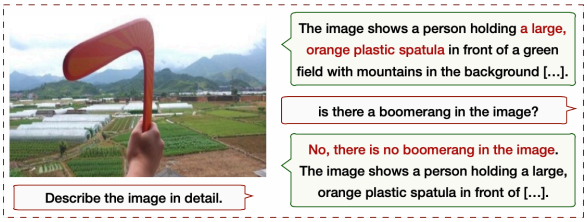
3.4 KG-aware Mutli-modal Pre-training

Currently, multi-modal LLMs (MLLMs) are widely recognized for their task-agnostic capabilities, but their lack of long-tail knowledge often leads to cross-modal hallucinations, causing inconsistencies and errors, as evidenced in Fig. 7. To mitigate this issue, KG-aware Multi-modal Pre-training offers a viable strategy for knowledge infusion.

Specifically, Med-VLP (Chen et al., 2022b) employs structured medical knowledge entities from UMLS KG (Bodenreider, 2004) as mediators to align image and text features (Li et al., 2021), using a whole-entity mask strategy (Sun et al., 2019a) instead of sub-word masking. It focuses the model's attention on crucial medical information across modalities, enabling medical VLMs to gain domain-specific knowledge for knowledge-aware representations in downstream tasks. Ye et al. (2023) introduce a DANCE dataset where commonsense KG triples are turned into natural language riddles paired with images, aimed at embedding knowledge relations and linking KG entries (h, r, t) with images that depict related entities, where entities in the images are referred to as "`this item`". KGTransformer (Zhang et al., 2023b) is a BERT-alike KG pre-training model with objectives like triple-based masked relation/entity prediction and entity pair prediction (i.e., assessing whether two entities from a one-hop subgraph were previously connected by the same relation in $\mathcal{G}$), supporting downstream tasks like QA and ZS-IMGC.



(*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 multi-modal LLMs (Zha et al., 2023), as demonstrated by (a) BLIP-2 (Li et al., 2023b) and (b) MiniGPT-4 (Zhu et al., 2023), which lead to multi-modal hallucinations.

4 Challenges and Future Directions

Multi-modal Knowledge Graphs. Focusing solely on the benefits of traditional KGs for multi-modal tasks can be limiting due to the narrow scope of knowledge in single-modality KGs. Scenarios like detailing badge designs or architectural photos, which are challenging to describe with text alone, highlight the necessity for MMKGs which integrate knowledge symbols across modalities like text, images, and sound (typically as entities). However, most MMKG research remains focused on tasks within the In-MMKG framework like Mulit-modal Knowledge Graph Completion (Zhao et al., 2022), with limited exploration into MMKG-driven tasks. We identify three main challenges hindering broader application: *non-uniform organization and ontology of MMKGs*, *substantial storage and processing overheads*, and issues of *data timeliness and completeness in MMKGs*, which are discussed in the Appendix. Currently, the few studies on MMKG-driven tasks primarily emphasize retrieval-related activities, exploiting MMKGs' natural database-like functionalities, yet they do not fully exploit MMKGs' structured multi-modal capabilities. For instance, Zha et al. (2023) enhance knowledge-based VQA by employing multi-modal concept descriptions and using MMKGs as "*key:value*" based retrieval KBs to support reasoning in MLLMs. Looking forward, to fully unlock the potential of large-scale MMKGs for multi-modal tasks, we must resolve key issues including the effective construction and utilization of MMKGs that are well-suited to multi-modal tasks.

Large Language Models. The challenge of extracting sufficient visual knowledge, as identified by Chen et al. (2023a), alongside Zhou et al.'s (2023) finding that 43% of BLIP2 (Li et al., 2023b) errors on the A-OKVQA dataset (Schwenk et al., 2022) could be addressed with proper knowledge integration. Fine-tuning MLLMs with MMKGs can be realized via two main strategies: active KG routing for creating specific instructions (Wan et al., 2023), and the use of self-instructing techniques to autonomously generate multi-grained, multi-modal instructional data (Du et al., 2023). Besides, the structured multi-modal relational data inherent in MMKGs provides an essential foundation for investigating the visual extrapolation abilities of Large Vision Models (LVMs) (Bai et al., 2023) as well as MLLMs (Sun et al., 2023d).

The increasing risk of generating seemingly authentic but factually inaccurate web content in MLLMs, known as *hallucination* (Agrawal et al., 2023), is another concern. Future efforts could focus on combining MMKGs with hallucination detection or correction methods to enhance multi-modal task precision and leveraging KGs for rewriting input statements in a knowledge-aware manner to reduce factual hallucinations in MLLM reasoning (Wang et al., 2023a). This process could be aligned with CoT approaches like Think-on-Graph (ToG) (Sun et al., 2023a), improving MLLM' abilities to interpret and interact with (MM)KGs, thus advancing towards human-like multi-modal proficiency and greater machine intelligence. Moreover, MMKGs can enhance multi-modal Retrieval Augmented Generation (RAG) by using various modalities as anchors (Song et al., 2023a), yielding more relevant and insightful results than traditional vector-based searches (Wu and Xie, 2023).

5 Conclusion and Vision

This paper provides an overview of KG-driven multi-modal learning, tracing the field's evolution from past achievements through current trends to future developments. Our goal is to construct a systematic blueprint of the domain, providing a valuable reference for researchers currently involved in or planning to explore this area. Looking ahead, we envision a stronger synergy between MMKGs and MLLMs, aiming to create a robust interactive system powered by this dual-drive approach.

6 Limitations

In this study, we provide a survey of multi-modal learning with knowledge graphs. 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. 1 and Tab. 2) 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. Additionally, considering the similarity and coupling between different tasks and the uneven number of related works, we may have overlooked some multi-modal tasks such as multi-modal summarization (Jangra et al., 2023). We also have not discussed specialized domains such as Medicine (Du et al., 2016) and Science, but we aim to address these gaps in the future.

Empirical Conclusions. We provide detailed comparisons and discussions on KG-driven multi-modal learning in § 3, listing some promising future directions in § 4. 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

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermüller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *CoRR*, abs/1511.03292.

Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 2023. Explicit knowledge integration for knowledge-aware visual question answering about named entities. In *ICMR*, pages 29–38. ACM.

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

Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. 2013. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826. IEEE Computer Society.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *ECCV (5)*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *ICCV*, pages 2425–2433. IEEE Computer Society.

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.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV (5)*, volume 12626 of *Lecture Notes in Computer Science*, pages 460–479. Springer.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING-ACL*, pages 86–90. Morgan Kaufmann Publishers / ACL.

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.

Hédi Ben-Younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: multimodal tucker fusion for visual question answering. In *ICCV*, pages 2631–2639. IEEE Computer Society.

Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. 2020. Do dogs have whiskers? A new knowledge base of haspart relations. *CoRR*, abs/2006.07510.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

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.

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.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Yuqi Bu, Xin Wu, Liuwu Li, Yi Cai, Qiong Liu, and Qingbao Huang. 2023. Segment-level and category-oriented network for knowledge-based referring expression comprehension. In *ACL (Findings)*, pages 8745–8757. Association for Computational Linguistics.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022a. Image-text retrieval: A survey on recent research and development. In *IJCAI*, pages 5410–5417. ijcai.org.

Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. 2022b. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Trans. Neural Networks Learn. Syst.*, 33(7):2758–2767.

Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. 2023. A comprehensive survey of scene graphs: Generation and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):1–26.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750. ACL.

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

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021a. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *AAAI*, pages 999–1008. AAAI Press.

Jiali Chen, Zhenjun Guo, Jiayuan Xie, Yi Cai, and Qing Li. 2023b. Deconfounded visual question generation with causal inference. In *ACM Multimedia*, pages 5132–5142. ACM.

Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z. Pan, and Huajun Chen. 2021b. Knowledge-aware zero-shot learning: Survey and perspective. In *IJCAI*, pages 4366–4373. ijcai.org.

Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Jeff Z. Pan, Yuan He, Wen Zhang, Ian Horrocks, and Huajun Chen. 2023c. Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. *Proc. IEEE*, 111(6):653–685.

Jiaoyan Chen, Freddy Lécué, Yuxia Geng, Jeff Z. Pan, and Huajun Chen. 2020a. Ontology-guided semantic composition for zero-shot learning. In *KR*, pages 850–854.

Jingjing Chen, Liangming Pan, Zhipeng Wei, Xiang Wang, Chong-Wah Ngo, and Tat-Seng Chua. 2020b. Zero-shot ingredient recognition by multi-relational graph convolutional network. In *AAAI*, pages 10542–10550. AAAI Press.

Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *CVPR*, pages 4042–4050. Computer Vision Foundation / IEEE Computer Society.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2023d. Large language models are visual reasoning coordinators. *CoRR*, abs/2310.15166.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter,

Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021c. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794. ACM.

Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 2019. Knowledge-embedded routing network for scene graph generation. In *CVPR*, pages 6163–6171. Computer Vision Foundation / IEEE.

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

Xiaolin Chen, Xuemeng Song, Liqiang Jing, Shuo Li, Linmei Hu, and Liqiang Nie. 2022a. Multimodal dialog systems with dual knowledge-enhanced generative pretrained language model. *CoRR*, abs/2207.07934.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020c. UNITER: universal image-text representation learning. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Zhanwen Chen, Saed Rezayi, and Sheng Li. 2023f. More knowledge, less bias: Unbiasing scene graph generation with explicit ontological adjustment. In *WACV*, pages 4012–4021. IEEE.

Zhihong Chen, Guanbin Li, and Xiang Wan. 2022b. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *ACM Multimedia*, pages 5152–5161. ACM.

Zhihong Chen, Ruifei Zhang, Yibing Song, Xiang Wan, and Guanbin Li. 2023g. Advancing visual grounding with scene knowledge: Benchmark and method. In *CVPR*, pages 15039–15049. IEEE.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. 2021d. Zero-shot visual question answering using knowledge graph. In *ISWC*, volume 12922 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. 2022c. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. In *IJCKG*, pages 20–29. ACM.

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

Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. 2020. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *CVPR*, pages 10908–10917. Computer Vision Foundation / IEEE.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *EMNLP*, pages 3829–3846. Association for Computational Linguistics.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*. ACM.

Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021. ROSITA: enhancing vision-and-language semantic alignments via cross- and intra-modal knowledge integration. In *ACM Multimedia*, pages 797–806. ACM.

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.

Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *CVPR*, pages 5079–5088. IEEE.

Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. 2024. Modality-aware integration with large language models for knowledge-based visual question answering. *Preprint*, arXiv:2402.12728.

Xiao Dong, Xunlin Zhan, Yunchao Wei, Xiaoyong Wei, Yaowei Wang, Minlong Lu, Xiaochun Cao, and Xiaodan Liang. 2023. Entity-graph enhanced cross-modal

pretraining for instance-level product retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13117–13133.

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

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.

Jiao Du, Weisheng Li, Ke Lu, and Bin Xiao. 2016. An overview of multi-modal medical image fusion. *Neurocomputing*, 215:3–20.

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

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. In *ICLR*. OpenReview.net.

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.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370. The Association for Computer Linguistics.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129.

Jingru Gan, Xinzhe Han, Shuhui Wang, and Qingming Huang. 2023. Open-set knowledge-based visual question answering with inference paths. *CoRR*, abs/2310.08148.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya N. Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *CVPR*, pages 5057–5067. IEEE.

Jingying Gao, Qi Wu, Alan Blair, and Maurice Pagnucco. 2023. Lora: A logical reasoning augmented dataset for visual question answering. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2019. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, pages 8303–8311. AAAI Press.

Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip H. S. Torr. 2021. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2):652–662.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Knowit VQA: answering knowledge-based questions about videos. In *AAAI*, pages 10826–10834. AAAI Press.

Diego García-Olano, Yasumasa Onoe, and Joydeep Ghosh. 2022. Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. In *WWW (Companion Volume)*, pages 705–715. ACM.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lécué. 2020. Conceptbert: Concept-aware representation for visual question answering. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 489–498. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640.

Ning Ge, Yonghua Zhu, Xiaoyu Xiong, Binghui Zheng, and Jieyu Huang. 2021. Knhigan: Knowledge-enhanced hierarchical generative adversarial network for fine-grained text-to-image synthesis. In *ISCID*, pages 357–360. IEEE.

Yuxia Geng, Jiaoyan Chen, Zhuo Chen, Jeff Z. Pan, Zhiquan Ye, Zonggang Yuan, Yantao Jia, and Huajun Chen. 2021a. Ontozsl: Ontology-enhanced zero-shot learning. In *WWW*, pages 3325–3336. ACM / IW3C2.

Yuxia Geng, Jiaoyan Chen, Zhiquan Ye, Zonggang Yuan, Wei Zhang, and Huajun Chen. 2021b. Explainable zero-shot learning via attentive graph convolutional network and knowledge graphs. *Semantic Web*, 12(5):741–765.

Yuxia Geng, Jiaoyan Chen, Wen Zhang, Yajing Xu, Zhuo Chen, Jeff Z. Pan, Yufeng Huang, Feiyu Xiong, and Huajun Chen. 2022. Disentangled ontology embedding for zero-shot learning. In *KDD*, pages 443–453. ACM.

Yuxia Geng, Jiaoyan Chen, Xiang Zhuang, Zhuo Chen, Jeff Z. Pan, Juan Li, Zonggang Yuan, and Huajun Chen. 2023. Benchmarking knowledge-driven zero-shot learning. *J. Web Semant.*, 75:100757.

Deepanway Ghosal, Navonil Majumder, Roy Ka-Wei Lee, Rada Mihalcea, and Soujanya Poria. 2023. Language guided visual question answering: Elevate your multimodal language model using knowledge-enriched prompts. *CoRR*, abs/2310.20159.

José Manuél Gómez-Pérez and Raúl Ortega. 2019. Look, read and enrich - learning from scientific figures and their captions. In *K-CAP*, pages 101–108. ACM.

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.

Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, pages 1969–1978. Computer Vision Foundation / IEEE.

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.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *NAACL-HLT*, pages 956–968. Association for Computational Linguistics.

Dan Guo, Hui Wang, and Meng Wang. 2022a. Context-aware graph inference with knowledge distillation for visual dialog. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6056–6073.

Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. 2020. Iterative context-aware graph inference for visual dialog. In *CVPR*, pages 10052–10061. Computer Vision Foundation / IEEE.

Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan S. Kankanhalli. 2022b. A unified end-to-end retriever-reader framework for knowledge-based VQA. In *ACM Multimedia*, pages 2061–2069. ACM.

Yudong Han, Jianhua Yin, Jianlong Wu, Yinwei Wei, and Liqiang Nie. 2023. Semantic-aware modular capsule routing for visual question answering. *IEEE Trans. Image Process.*, 32:5537–5549.

Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuan-Fang Li. 2020. Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation. In *IJCAI*, pages 587–593. ijcai.org.

Xuehai He and Xin Wang. 2023. Multimodal graph transformer for multimodal question answering. In *EACL*, pages 189–200. Association for Computational Linguistics.

Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi, and Byoung-Tak Zhang. 2022. Hypergraph transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. In *ACL (1)*, pages 373–390. Association for Computational Linguistics.

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

Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6):118:1–118:36.

Jingyi Hou, Xinxiao Wu, Yayun Qi, Wentian Zhao, Jiebo Luo, and Yunde Jia. 2019. Relational reasoning using prior knowledge for visual captioning. *CoRR*, abs/1906.01290.

Jingyi Hou, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo. 2020. Joint commonsense and relation reasoning for image and video captioning. In *AAAI*, pages 10973–10980. AAAI Press.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Kenneth Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual storytelling. In *AAAI*, pages 7952–7960. AAAI Press.

Chi-Yang Hsu, Yun-Wei Chu, Ting-Hao (Kenneth) Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4443–4453. Association for Computational Linguistics.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *CoRR*, abs/2211.09699.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *CVPR*, pages 23369–23379. IEEE.

Feicheng Huang, Zhixin Li, Shengjia Chen, Canlong Zhang, and Huifang Ma. 2020a. Image captioning with internal and external knowledge. In *CIKM*, pages 535–544. ACM.

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.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020b. Movienet: A holistic dataset for movie understanding. In *ECCV (4)*, volume 12349 of *Lecture Notes in Computer Science*, pages 709–727. Springer.

Yan Huang, Yuming Wang, Yunan Zeng, and Liang Wang. 2022. MACK: multimodal aligned conceptual knowledge for unpaired image-text matching. In *NeurIPS*.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). In *NeurIPS*, pages 10944–10956.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. 2023b. Structure-clip: Enhance multi-modal language representations with structure knowledge. *CoRR*, abs/2305.06152.

Afzaal Hussain, Ifrah Maqsood, Muhammad Shahzad, and Muhammad Moazam Fraz. 2022. Multimodal knowledge reasoning for enhanced visual question answering. In *SITIS*, pages 224–230. IEEE.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, pages 6384–6392. AAAI Press.

Filip Ilievski, Pedro A. Szekely, and Bin Zhang. 2021. CSKG: the commonsense knowledge graph. In *ESWC*, volume 12731 of *Lecture Notes in Computer Science*, pages 680–696. Springer.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*, pages 874–880. Association for Computational Linguistics.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. 2021. Perceiver: General perception with iterative attention. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *SIGIR*, pages 2491–2498. ACM.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Comput. Surv.*, 55(13s):296:1–296:36.

Xiaoze Jiang, Siyi Du, Zengchang Qin, Yajing Sun, and Jing Yu. 2020. KBGN: knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue. In *ACM Multimedia*, pages 1265–1273. 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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

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.

Jun Cheng and Fuxiang Wu and Yanling Tian and Lei Wang and Dapeng Tao. 2022. Rifegan2: Rich feature generation for text-to-image synthesis from constrained prior knowledge. *IEEE Trans. Circuits Syst. Video Technol.*, 32(8):5187–5200.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *ACL (1)*, pages 5591–5606. Association for Computational Linguistics.

Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. 2019. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, pages 11487–11496. Computer Vision Foundation / IEEE.

Gi-Cheon Kang, Junseok Park, Hwaran Lee, Byoung-Tak Zhang, and Jin-Hwa Kim. 2021. Reasoning visual dialog with sparse graph learning and knowledge transfer. In *EMNLP (Findings)*, pages 327–339. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781. Association for Computational Linguistics.

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. Mm-reasoner: A multimodal knowledge-aware framework for knowledge-based visual question answering. In *EMNLP (Findings)*, pages 6571–6581. Association for Computational Linguistics.

Muhammad Jaleed Khan, John G. Breslin, and Edward Curry. 2022a. Common sense knowledge infusion for visual understanding and reasoning: Approaches, challenges, and applications. *IEEE Internet Comput.*, 26(4):21–27.

Muhammad Jaleed Khan, John G. Breslin, and Edward Curry. 2022b. Expressive scene graph generation using commonsense knowledge infusion for visual understanding and reasoning. In *ESWC*, volume 13261 of *Lecture Notes in Computer Science*, pages 93–112. Springer.

Apoorv Khandelwal, Ellie Pavlick, and Chen Sun. 2023. Analyzing modular approaches for visual question decomposition. *CoRR*, abs/2311.06411.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NeurIPS*, pages 1571–1581.

Barbara Ann Kipfer. 1992. Roget's 21st century thesaurus in dictionary form: the essential reference for home, school, or office. *(No Title)*.

Rajat Koner, Hang Li, Marcel Hildebrandt, Deepan Das, Volker Tresp, and Stephan Günnemann. 2021. Graphhopper: Multi-hop scene graph reasoning for visual question answering. In *ISWC*, volume 12922 of *Lecture Notes in Computer Science*, pages 111–127. Springer.

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.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1378–1387. JMLR.org.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*. OpenReview.net.

Quoc V. Le and Tomás Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Multi-label zero-shot learning with structured knowledge graphs. In *CVPR*, pages 1576–1585. Computer Vision Foundation / IEEE Computer Society.

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.

Adam Lerer, Ledell Wu, Jiajun Shen, Timothée Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. Pytorch-biggraph: A large scale graph embedding system. In *MLSys*. mlsys.org.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G. Moreno, and Jesús Lovón-Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *SIGIR*, pages 3108–3120. ACM.

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.

Guohao Li, Hang Su, and Wenwu Zhu. 2017. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. *CoRR*, abs/1712.00733.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *ACM Multimedia*, pages 1227–1235. ACM.

Jiaqi Li, Guilin Qi, Chuanyi Zhang, Yongrui Chen, Yiming Tan, Chenlong Xia, and Ye Tian. 2023a. Incorporating domain knowledge graph into multimodal movie genre classification with self-supervised attention and contrastive learning. In *ACM Multimedia*, pages 3337–3345. ACM.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. 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.

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.

Mingxiao Li and Marie-Francine Moens. 2022. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. In *AAAI*, pages 10983–10992. AAAI Press.

Rengang Li, Cong Xu, Zhenhua Guo, Baoyu Fan, Runze Zhang, Wei Liu, Yaqian Zhao, Weifeng Gong, and Endong Wang. 2022b. AI-VQA: visual question answering based on agent interaction with interpretability. In *ACM Multimedia*, pages 5274–5282. ACM.

Tengpeng Li, Hanli Wang, Bin He, and Chang Wen Chen. 2023c. Knowledge-enriched attention network with group-wise semantic for visual storytelling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8634–8645.

Wenhui Li, Song Yang, Qiang Li, Xuanya Li, and An-An Liu. 2023d. Commonsense-guided semantic and relational consistencies for image-text retrieval. *IEEE Transactions on Multimedia*.

Xiangyang Li and Shuqiang Jiang. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Trans. Multim.*, 21(8):2117–2130.

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

Xirong Li, Shuai Liao, Weiyu Lan, Xiaoyong Du, and Gang Yang. 2015. Zero-shot image tagging by hierarchical semantic embedding. In *SIGIR*, pages 879–882. ACM.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. 2023f. A comprehensive evaluation of GPT-4V on knowledge-intensive visual question answering. *CoRR*, abs/2311.07536.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *EMNLP*, pages 11238–11254. Association for Computational Linguistics.

Weizhe Lin, Zhilin Wang, and Bill Byrne. 2023. FVQA 2.0: Introducing adversarial samples into fact-based visual question answering. In *EACL (Findings)*, pages 149–157. Association for Computational Linguistics.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. REVIVE: regional visual representation matters in knowledge-based visual question answering. In *NeurIPS*.

An-An Liu, Chenxi Huang, Ning Xu, Hongshuo Tian, Jing Liu, and Yongdong Zhang. 2023a. Counterfactual visual dialog: Robust commonsense knowledge learning from unbiased training. *IEEE Transactions on Multimedia*.

An-An Liu, Zefang Sun, Ning Xu, Rongbao Kang, Jinbo Cao, Fan Yang, Weijun Qin, Shenyuan Zhang, Jiaqi Zhang, and Xuanya Li. 2023b. Prior knowledge guided text to image generation. *Pattern Recognition Letters*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *CoRR*, abs/2304.08485.

Luping Liu, Meiling Wang, Xiaohai He, Linbo Qing, and Honggang Chen. 2022. Fact-based visual question answering via dual-process system. *Knowl. Based Syst.*, 237:107650.

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.

Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018a. Entity-aware image caption generation. In *EMNLP*, pages 4013–4023. Association for Computational Linguistics.

Jiale Lu, Lianggangxu Chen, Youqi Song, Shaohui Lin, Changbo Wang, and Gaoqi He. 2023. Prior knowledge-driven dynamic scene graph generation with causal inference. In *ACM Multimedia*, pages 4877–4885. ACM.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10434–10443. Computer Vision Foundation / IEEE.

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.

Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018b. R-VQA: learning visual relation facts with semantic attention for visual question answering. In *KDD*, pages 1880–1889. ACM.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *EMNLP (1)*, pages 6417–6431. 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.

Maria Lymperaiou and Giorgos Stamou. 2022. A survey on knowledge-enhanced multimodal learning. *CoRR*, abs/2211.12328.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *AAAI*, pages 2925–2933. AAAI Press.

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.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. KRISP: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *CVPR*, pages 14111–14121. Computer Vision Foundation / IEEE.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204. Computer Vision Foundation / IEEE.

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.

Aditya Mogadala, Xiaoyu Shen, and Dietrich Klakow. 2020. Integrating image captioning with rule-based entity masking. *CoRR*, abs/2007.11690.

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

Sebastian Monka, Lavdim Halilaj, and Achim Rettinger. 2022. A survey on visual transfer learning using knowledge graphs. *Semantic Web*, 13(3):477–510.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Jan Svoboda, and Michael M. Bronstein. 2017. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*, pages 5425–5434. IEEE Computer Society.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *ACL (1)*. The Association for Computer Linguistics.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *NeurIPS*, pages 2659–2670.

Medhini Narasimhan and Alexander G. Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *ECCV (8)*, volume 11212 of *Lecture Notes in Computer Science*, pages 460–477. Springer.

Nihal V. Nayak and Stephen H. Bach. 2022. Zero-shot learning with common sense knowledge graphs. *Trans. Mach. Learn. Res.*, 2022.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *WWW*, pages 2636–2647. ACM / IW3C2.

Weizhi Nie, Ruidong Chen, Weijie Wang, Bruno Lepri, and Nicu Sebe. 2023. T2TD: text-3d generation model based on prior knowledge guidance. *CoRR*, abs/2305.15753.

Sofia Nikiforova, Tejaswini Deoskar, Denis Paperno, and Yoad Winter. 2022. Generating image captions with external encyclopedic knowledge. *CoRR*, abs/2210.04806.

Timothy Ossowski and Junjie Hu. 2023. Multimodal prompt retrieval for generative visual question answering. *CoRR*, abs/2306.17675.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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.

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.

Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. 2023. Frozen transformers in language models are effective visual encoder layers. *CoRR*, abs/2310.12973.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

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.

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP/IJCNLP (1)*, pages 2463–2473. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *ACL (demo)*, pages 101–108. Association for Computational Linguistics.

Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Trans. Multim. Comput. Commun. Appl.*, 17(3):98:1–98:23.

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.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Learn, imagine and create: Text-to-image generation from prior knowledge. In *NeurIPS*, pages 885–895.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2021. Referring expression comprehension: A survey of methods and datasets. *IEEE Trans. Multim.*, 23:4426–4440.

Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik G. Learned-Miller. 2021. Passage retrieval for outside-knowledge visual question answering. In *SIGIR*, pages 1753–1757. ACM.

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.

Kiran Ramnath and Mark Hasegawa-Johnson and. 2020. Seeing is knowing! fact-based visual question answering using knowledge graph embeddings. *CoRR*, abs/2012.15484.

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2023. VLC-BERT: visual question answering with contextualized commonsense knowledge. In *WACV*, pages 1155–1165. IEEE.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767.

Benjamin Z. Reichman, Anirudh Sundar, Christopher Richardson, Tamara Zubatiy, Prithwijit Chowdhury, Aaryan Shah, Jack Truxal, Micah Grimes, Dristi Shah, Woo Ju Chee, Saif Punjwani, Atishay Jain, and Larry Heck. 2023. Outside knowledge visual question answering version 2.0. In *ICASSP*, pages 1–5. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics.

Abhinaba Roy, Deepanway Ghosal, Erik Cambria, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. Improving zero-shot learning baselines with commonsense knowledge. *Cogn. Comput.*, 14(6):2212–2222.

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.

Tara Safavi and Danai Koutra. 2020. Codex: A comprehensive knowledge graph completion benchmark. In *EMNLP (1)*, pages 8328–8350. Association for Computational Linguistics.

Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. 2023. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Syst. Appl.*, 212:118669.

Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023a. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In *SIGIR*, pages 110–120. ACM.

Alireza Salemi, Mahta Rafiee, and Hamed Zamani. 2023b. Pre-training multi-modal dense retrievers for outside-knowledge visual question answering. In *ICTIR*, pages 169–176. ACM.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *NIPS*, pages 2226–2234.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, pages 3027–3035. AAAI Press.

Raeid Saqur and Karthik Narasimhan. 2020. Multimodal graph networks for compositional generalization in visual question answering. In *NeurIPS*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *CoRR*, abs/2304.15004.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *ESWC*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

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.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: knowledge-aware visual question answering. In *AAAI*, pages 8876–8884. AAAI Press.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *CVPR*, pages 14974–14983. IEEE.

Violetta Shevchenko, Damien Teney, Anthony R. Dick, and Anton van den Hengel. 2021. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *CoRR*, abs/2101.06013.

Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. 2019. Knowledge aware semantic concept expansion for image-text matching. In *IJCAI*, pages 5182–5189. ijcai.org.

Zhan Shi, Yilin Shen, Hongxia Jin, and Xiaodan Zhu. 2022. Improving zero-shot phrase grounding via reasoning on external knowledge and spatial relations. In *AAAI*, pages 2253–2261. AAAI Press.

Qingyi Si, Yuchen Mo, Zheng Lin, Huishan Ji, and Weiping Wang. 2023. Combo of thinking and observing for outside-knowledge VQA. In *ACL (1)*, pages 10959–10975. Association for Computational Linguistics.

Sibei Yang and Guanbin Li and Yizhou Yu. 2021. Relationship-embedded representation learning for grounding referring expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(8):2765–2779.

Linda B. Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artif. Life*, 11(1-2):13–29.

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

Lingyun Song, Jianao Li, Jun Liu, Yang Yang, Xuequn Shang, and Mingxuan Sun. 2023b. Answering knowledge-based visual questions via the exploration of question purpose. *Pattern Recognit.*, 133:109015.

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*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR*, pages 2443–2449. ACM.

Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. 2018. Learning visual knowledge memory networks for visual question answering. In *CVPR*, pages 7736–7745. Computer Vision Foundation / IEEE Computer Society.

Sanjay Subramanian, Medhini Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. 2023. Modular visual question answering via code generation. In *ACL (2)*, pages 747–761. Association for Computational Linguistics.

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

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023a. 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. 2023b. Head-to-tail: How knowledgeable are large language models (llm)? A.K.A. will llms replace knowledge graphs? *CoRR*, abs/2308.10168.

Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and Philip S. Yu. 2023c. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Trans. Knowl. Data Eng.*, 35(12):12736–12749.

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.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019a. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

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

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics.

Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. 2023. Knowledge-based embodied question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11948–11960.

Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. Webchild: harvesting and organizing commonsense knowledge from the web. In *WSDM*, pages 523–532. ACM.

Wei Tang, Liang Li, Xuejing Liu, Lu Jin, Jinhui Tang, and Zechao Li. 2023. Context disentangling and prototype inheriting for robust visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hongshuo Tian, Ning Xu, Yanhui Wang, Chenggang Yan, Bolun Zheng, Xuanya Li, and An-An Liu. 2023. Towards confidence-aware commonsense knowledge integration for scene graph generation. In *ICME*, pages 2255–2260. IEEE.

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.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *CVSC*, pages 57–66. Association for Computational Linguistics.

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.

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.

Kohei Uehara and Tatsuya Harada. 2023. K-VQG: knowledge-aware visual question generation for common-sense acquisition. In *WACV*, pages 4390–4398. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society.

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

Peter Vickers, Nikolaos Aletras, Emilio Monti, and Loïc Barrault. 2021. In factuality: Efficient integration of relevant facts for visual question answering. In *ACL/IJCNLP (2)*, pages 468–475. Association for Computational Linguistics.

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

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.

Hai Wan, Yonghao Luo, Bo Peng, and Wei-Shi Zheng. 2018. Representation learning for scene graph completion via jointly structural and visual embedding. In *IJCAI*, pages 949–956. ijcai.org.

Haoran Wang, Dongliang He, Wenhao Wu, Boyang Xia, Min Yang, Fu Li, Yunlong Yu, Zhong Ji, Errui Ding, and Jingdong Wang. 2022a. CODER: coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In *ECCV (36)*, volume 13696 of *Lecture Notes in Computer Science*, pages 700–716. Springer.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020a. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV (24)*, volume 12369 of *Lecture Notes in Computer Science*, pages 18–34. Springer.

Huan Wang, Weiming Lu, and Zeyun Tang. 2019. Incorporating external knowledge to boost machine comprehension based question answering. In *ECIR (1)*, volume 11437 of *Lecture Notes in Computer Science*, pages 819–827. Springer.

Jin Wang and Bo Jiang. 2021. Zero-shot learning via contrastive learning on dual knowledge graphs. In *ICCVW*, pages 885–892. IEEE.

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

Peng Wang, Dongyang Liu, Hui Li, and Qi Wu. 2020b. Give me something to eat: Referring expression comprehension with commonsense knowledge. In *ACM Multimedia*, pages 28–36. ACM.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, pages 1290–1296. ijcai.org.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2018a. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427.

Xiaodan Wang, Lei Li, Zhixu Li, Xuwu Wang, Xiangru Zhu, Chengyu Wang, Jun Huang, and Yanghua Xiao. 2023b. AGREE: aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models. In *WSDM*, pages 456–464. ACM.

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018b. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866. Computer Vision Foundation / IEEE Computer Society.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. 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. 2023c. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *ACM Multimedia*, pages 2391–2399. ACM.

Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. 2022b. VQA-GNN: reasoning with multimodal semantic graph for visual question answering. *CoRR*, abs/2205.11501.

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. 2020c. Deep multimodal fusion by channel exchanging. In *NeurIPS*.

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

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020d. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *ICMR*, pages 540–547. ACM.

Zeqing Wang, Wentao Wan, Runmeng Chen, Qiqing Lao, Minjie Lang, and Keze Wang. 2023e. 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. 2023f. Learning to filter context for retrieval-augmented generation. *CoRR*, abs/2311.08377.

Zihao Wang, Junli Wang, and Changjun Jiang. 2022c. Unified multimodal model with unlikelihood training for visual dialog. In *ACM Multimedia*, pages 4625–4634. ACM.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. 2023g. Filling the image information gap for VQA: prompting large language models to proactively ask questions. *CoRR*, abs/2311.11598.

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*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based VQA. In *AAAI*, pages 2712–2721. AAAI Press.

Jialin Wu and Raymond J. Mooney. 2022. Entity-focused dense passage retrieval for outside-knowledge visual question answering. In *EMNLP*, pages 8061–8072. Association for Computational Linguistics.

Likang Wu, Zhi Li, Hongke Zhao, Zhefeng Wang, Qi Liu, Baoxing Huai, Nicholas Jing Yuan, and Enhong Chen. 2023a. Recognizing unseen objects via multimodal intensive knowledge graph propagation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 2618–2628. ACM.

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

Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, pages 4622–4630. IEEE Computer Society.

Sen Wu, Guoshuai Zhao, and Xueming Qian. 2023b. Resolving zero-shot and fact-based visual question answering via enhanced fact retrieval. *IEEE Transactions on Multimedia*.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *SIGMOD Conference*, pages 481–492. ACM.

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

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023. A simple baseline for knowledge-based visual question answering. *CoRR*, abs/2310.13570.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2251–2265.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551. Computer Vision Foundation / IEEE Computer Society.

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.

Jiayuan Xie, Wenhao Fang, Yi Cai, Qingbao Huang, and Qing Li. 2022. Knowledge-based visual question generation. *IEEE Trans. Circuits Syst. Video Technol.*, 32(11):7547–7558.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. KM-BART: knowledge enhanced multimodal BART for visual commonsense generation. In *ACL/IJCNLP (1)*, pages 525–535. Association for Computational Linguistics.

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

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *CoRR*, abs/1304.5634.

Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. 2021. Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning. In *AAAI*, pages 3022–3029. AAAI Press.

Da Xu, Chuanwei Ruan, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. 2020. Product knowledge graph embedding for e-commerce. In *WSDM*, pages 672–680. ACM.

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

Silei Xu, Shicheng Liu, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu, Sina J. Semnani, and Monica S. Lam. 2023c. 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.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324. Computer Vision Foundation / IEEE Computer Society.

Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019a. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *IJCAI*, pages 5356–5362. ijcai.org.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2019b. Cross-modal relationship inference for grounding referring expressions. In *CVPR*, pages 4145–4154. Computer Vision Foundation / IEEE.

Song Yang, Qiang Li, Wenhui Li, Min Liu, Xuanya Li, and Anan Liu. 2023a. External knowledge dynamic modeling for image-text retrieval. In *ACM Multimedia*, pages 5330–5338. ACM.

Xiao Yang, Xiaojun Wu, and Tianyang Xu. 2021. DRSGN: dual revised semantic graph structured network for image-text matching. In *CCIS*, pages 230–242. IEEE.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *AAAI*, pages 3081–3089. AAAI Press.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *CVPR*, pages 21–29. IEEE Computer Society.

Zonglin Yang, Xinya Du, Erik Cambria, and Claire Cardie. 2023b. End-to-end case-based reasoning for commonsense knowledge base completion. In *EACL*, pages 3491–3504. Association for Computational Linguistics.

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

Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. 2023. Improving commonsense in vision-language models via knowledge graph riddles. In *CVPR*, pages 2634–2645. IEEE.

Chengxiang Yin, Zhengping Che, Kun Wu, Zhiyuan Xu, and Jian Tang. 2023a. Multi-clue reasoning with memory augmentation for knowledge-based visual question answering. *CoRR*, abs/2312.12723.

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. 2023b. 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.

Jiuxiang You, Zhenguo Yang, Qing Li, and Wenyin Liu. 2023. A retriever-reader framework with visual entity linking for knowledge-based visual question answering. In *ICME*, pages 13–18. IEEE.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI*, pages 3208–3216. AAAI Press.

Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognit.*, 108:107563.

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.

Xiang Yu, Ruoxin Chen, Jie Li, Jiawei Sun, Shijing Yuan, Huxiao Ji, Xinyu Lu, and Chentao Wu. 2022. Zero-shot scene graph generation with knowledge graph completion. In *ICME*, pages 1–6. IEEE.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. 2020a. Bridging knowledge graphs to generate scene graphs. In *ECCV (23)*, volume 12368 of *Lecture Notes in Computer Science*, pages 606–623. Springer.

Alireza Zareian, Zhecan Wang, Haoxuan You, and Shih-Fu Chang. 2020b. Learning visual commonsense for robust scene graph generation. In *ECCV (23)*, volume 12368 of *Lecture Notes in Computer Science*, pages 642–657. Springer.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731. Computer Vision Foundation / IEEE.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840. Computer Vision Foundation / IEEE Computer Society.

Zeynep Akata and Florent Perronnin and Zaïd Harchaoui and Cordelia Schmid. 2016. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1425–1438.

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.

Chenrui Zhang, Xiaoqing Lyu, and Zhi Tang. 2019. TGG: transferable graph generation for zero-shot and few-shot learning. In *ACM Multimedia*, pages 1641–1649. ACM.

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.

Jiawen Zhang, Abhijit Mishra, Avinesh P. V. S., Siddharth Patwardhan, and Sachin Agarwal. 2022a. Can open domain question answering systems answer visual knowledge questions? *CoRR*, abs/2202.04306.

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

Liyang Zhang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. 2021a. Rich visual knowledge-based augmentation network for visual question answering. *IEEE Trans. Neural Networks Learn. Syst.*, 32(10):4362–4373.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588. Computer Vision Foundation / IEEE.

Shunyu Zhang, Xiaoze Jiang, Zequn Yang, Tao Wan, and Zengchang Qin. 2022b. Reasoning with multi-structure commonsense knowledge in visual dialog. In *CVPR Workshops*, pages 4599–4608. IEEE.

Wen Zhang, Yushan Zhu, Mingyang Chen, Yuxia Geng, Yufeng Huang, Yajing Xu, Wenting Song, and Huajun Chen. 2023b. Structure pretraining and prompt tuning for knowledge graph transfer. In *WWW*, pages 2581–2590. ACM.

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

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *AAAI*, pages 12910–12917. AAAI Press.

Yu Zhang, Xinyu Shi, Siya Mi, and Xu Yang. 2021c. Image captioning with transformer and knowledge graph. *Pattern Recognit. Lett.*, 143:43–49.

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.

Yuchen Zhang, Xing Su, Jia Wu, Jian Yang, Hao Fan, and Xiaochuan Zheng. 2023d. Emoknow: Emotion- and knowledge-oriented model for COVID-19 fake news detection. In *ADMA (1)*, volume 14176 of *Lecture Notes in Computer Science*, pages 352–367. Springer.

Yuhong Zhang, Haitao Shu, Chenyang Bu, and Xuegang Hu. 2022d. A zero-shot learning method with a multi-modal knowledge graph. In *ICTAI*, pages 391–395. IEEE.

Zefan Zhang, Yi Ji, and Chunping Liu. 2023e. Knowledge-aware causal inference network for visual dialog. In *ICMR*, pages 253–261. ACM.

Zhi Zhang, Helen Yannakoudakis, Xiantong Zhen, and Ekaterina Shutova. 2023f. Ck-transformer: Commonsense knowledge enhanced transformers for referring expression comprehension. In *EACL (Findings)*, pages 2541–2551. Association for Computational Linguistics.

Lei Zhao, Lianli Gao, Yuyu Guo, Jingkuan Song, and Heng Tao Shen. 2021a. Skanet: Structured knowledge-aware network for visual dialog. In *ICME*, pages 1–6. IEEE.

Lei Zhao, Junlin Li, Lianli Gao, Yunbo Rao, Jingkuan Song, and Heng Tao Shen. 2023. Heterogeneous knowledge network for visual dialog. *IEEE Trans. Circuits Syst. Video Technol.*, 33(2):861–871.

Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo. 2021b. 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. 2022. Mose: Modality split and ensemble for multimodal knowledge graph completion. In *EMNLP*, pages 10527–10536. Association for Computational Linguistics.

Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. 2021a. Knowledge is power: Hierarchical-knowledge embedded meta-learning for visual reasoning in artistic domains. In *KDD*, pages 2360–2368. ACM.

Wenfeng Zheng, Lirong Yin, Xiaobing Chen, Zhiyang Ma, Shan Liu, and Bo Yang. 2021b. Knowledge base graph embedding module design for visual question answering model. *Pattern Recognit.*, 120:108153.

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.

Jiaguo Zhong and Dongsheng Wang. 2023. Image caption generation based on object detection and knowledge enhancement. In *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023)*, volume 12707, pages 226–232. SPIE.

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.

Yimin Zhou, Yiwei Sun, and Vasant G. Honavar. 2019. Improving image captioning by leveraging knowledge graphs. In *WACV*, pages 283–293. IEEE.

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

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

Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, pages 1004–1013. Computer Vision Foundation / IEEE Computer Society.

Yonghua Zhu, Ning Ge, Jieyu Huang, Yunwen Zhu, Binghui Zheng, and Wenjun Zhang. 2021. Enriching attributes from knowledge graph for fine-grained text-to-image synthesis. In *CSAE*, pages 80:1–80:6. ACM.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004. IEEE Computer Society.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *IJCAI*, pages 1097–1103. ijcai.org.

Maryam Ziaeefard and Freddy Lécué. 2020. Towards knowledge-augmented visual question answering. In *COLING*, pages 1863–1873. International Committee on Computational Linguistics.

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. We begin by introducing the preliminaries, defining key concepts in KGs and multi-modal learning, and providing an overview of KG4MML settings (§ 2). Then we delves into various KG4MM tasks, detailing resources and key breakthroughs in recent years (§ 3), while balancing details to address content overlaps across tasks and focusing on core challenges (Fig. 2). Finally, we explore the future integration of multi-modal methods with (MM)KGs, proposing potential enhancements for previously discussed tasks (§ 4), especially considering the rapid development of multi-modal Large Language Models (LLMs).

Related work. Several surveys are closely related to our work. Monka et al. (2022) overview Knowledge Graph Embedding (KGE) methods and their integration with high-dimensional visual embeddings, emphasizing KGs' role in visual information transfer. Lymperaiou and Stamou (2022) discuss enhancing multi-modal learning with knowledge but lack a systematic analysis of KG-driven multi-modal learning, including benchmarks, method comparisons, and task paradigms. Moreover, these studies all focus on developments up to 2022, missing the latest insights.

In response to the rapid advancements in AGI from 2022 to 2023, our survey places a strong emphasis on emerging areas like LLMs, aiming to fill critical knowledge gaps. Our goal is to provide a clear roadmap for future research, highlight challenges and opportunities, and systematically compare methodologies to inspire new ideas.

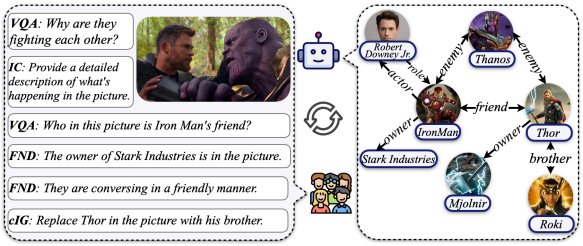


Figure 8: Applications of KGs in downstream multi-modal tasks within the context of the *Marvel Universe*.

A.2 Supplement for 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, and conclude with diverse interpretations and uses of KGs beyond the semantic web.

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.

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.

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., 2023b; Pan et al., 2023a,b).

Ontology. Within the semantic web, ontologies serve as KG schemas, utilizing languages like RDFS[6] and OWL[7] to ensure richer semantics and superior quality (Horrocks, 2008). Key features of ontologies include: ***(i)*** Hierarchical classes, often termed as concepts[8]; ***(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.

Languages like RDF, RDFS, and OWL introduce built-in vocabularies to capture these knowledge elements, with predicates like *rdfs:subClassOf* denoting concept subsumption, and *rdf:type* indicating instance-concept associations. RDFS also provides annotation properties like *rdfs:label* and *rdfs:comment* for resource meta-information.

KG Scope Extension. Widely accepted KGs include WordNet (Miller, 1995), a lexical database defining word interrelations, and ConceptNet (Speer et al., 2017), which archives commonsense knowledge interlinked by different terms.

In this paper, we extend the conventional view of KGs beyond standard-format entities and relations. This paper extends the conventional view of KGs beyond standard-format entities and relations. Besides, the ontology alone, often utilized to define domain knowledge including conceptualization and vocabularies like terms and taxonomies, is also considered a form of KG. Further elaborating on this expanded perspective, as outlined by Chen et al. (2023c), our scope includes simpler graph structures, such as basic taxonomies with hierarchical classes and graphs with weighted edges denoting quantitative relationships like similarity and distance between entities. Additionally, we categorize any structured data organized in a graph format with nodes that have explicit semantic interpretations as part of this broader KG definition. A prominent example is the Semantic Network, which connects various concepts with labeled edges to represent different relationships.

We also consider ontologies, basic taxonomies, and graphs with weighted edges as forms of KGs. Any structured data organized in a graph format with nodes having explicit semantic interpretations falls under this broader KG definition. An example is the Semantic Network, which connects various concepts with labeled edges to represent different relationships.

A.2.2 Multi-modal Learning.

Our world is perceived through diverse modalities, including sight, sound, movement, touch, and smell (Smith and Gasser, 2005). A "modality" typically refers to a specific type of data or information channel, characterized by sensory input or representation format. Each modality encapsulates unique features from specific sensory sources. It is intuitive that models, which integrate data from various modalities, generally surpass uni-modal models by accumulating more information. Multi-modal learning aims to develop a unified representation or mapping from multiple modalities to an output space, leveraging the complementarity and redundancy across modalities to improve prediction. The challenge lies in effectively aligning, fusing, and integrating information from various modalities to exploit their collective power.

Difference to Multi-view Learning. Unlike multi-view analysis, which suggests that each view

[4] https://lookup.dbpedia.org/

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

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

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

[8] To distinguish between *class* in machine learning tasks and *class* in KGs, we refer to the latter as *concept*.

(e.g., different perspectives of a flower) can independently yield accurate predictions (Xu et al., 2013; Federici et al., 2020), multi-modal learning contends with scenarios where the absence of one modality could impede task completion (Huang et al., 2021) (e.g., an image-lacking Visual Question Answering scenario). Additionally, multi-view learning typically involves varying perspectives of the same data type, originating from a single source, such as different features of image data. In contrast, multi-modal learning deals with disparate data types, like text and images, derived from multiple sources. In this paper, our exploration of multi-modal tasks and the application of multi-modal learning on KGs are grounded in this broader understanding of multi-modal learning.

Definition 1 ***Multi-modal Learning.*** *Assume given data* $\hat{x} = \left(x^{(1)}, \ldots, x^{(K)}\right)$ *consists of* K *modalities, with* $x^{(k)} \in \mathcal{X}^{(k)}$ *representing the domain set of the* k*-th modality and* $\mathcal{X} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(K)}$*. Let* $\mathcal{Y}$ *denote the target domain and* $\mathcal{Z}$ *represent a latent space. Denote* $g : \mathcal{X} \mapsto \mathcal{Z}$ *as the true mapping from the input space (utilizing all* K *modalities) to the latent space, and* $q : \mathcal{Z} \mapsto \mathcal{Y}$ *as the true task mapping. For example, in aggregation-based multi-modal fusion,* g *serves as an aggregation function built upon* K *separate sub-networks, and* q *is a multi-layer neural network (Wang et al., 2020c). In a learning task, a data pair* $(\hat{x}, y) \in \mathcal{X} \times \mathcal{Y}$ *is generated from an unknown distribution* $\mathcal{D}$*, such that*

$$\mathbb{P}_{\mathcal{D}}(\hat{x}, y) = \mathbb{P}_{y|\hat{x}}\left(y \mid q \circ g(\hat{x})\right) \mathbb{P}_{\hat{x}}(\hat{x}), \quad (1)$$

where $q \circ g(\hat{x}) = q(g(\hat{x}))$ *represents the composite function of* q *and* g*.*

A.3 Supplement for Understanding & Reasoning Tasks

Multi-modal reasoning tasks, like knowledge-based VQA, demand knowledge that goes beyond regular daily experiences (Khan et al., 2022a). These tasks often delve into less common, long-tail knowledge domains that typically require intentional reflection, with KGs providing a crucial structured repository for this extensive, specialized knowledge.

A.3.1 Supplementary Information for VQA

Current KG-aware VQA research typically involves four key stages for incorporating knowledge. These stages, integral to the workflow of

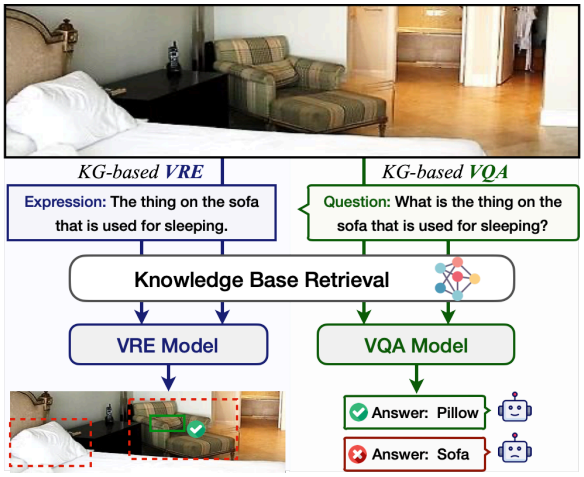


Figure 9: Illustration of KG-based Visual Question Answering (VQA) (§ 3.1) and Visual Referring Expressions (VRE) (§ A.6). To some extent, KG-based VRE can be viewed as an extension of KG-based VQA, incorporating an additional step of grounding answers.

KG-aware understanding and reasoning tasks, may be adopted individually or in combination across different studies to form a comprehensive approach. Those method references are provided and organized here for tracing the original sources:

Knowledge Retrieval. Implicit knowledge encoded in model parameters, typically pre-trained on large-scale datasets via self-supervised tasks, makes the use of a retriever **optional but still beneficial** for VQA.

1. Matching-based Retrieval:
   - Identifying spatial positions (Shah et al., 2019; Zhu et al., 2020; Yu et al., 2020; Gardères et al., 2020; Marino et al., 2021; Lin et al., 2022).
   - Identifying visual object sizes and names (Wang et al., 2017, 2018a; Narasimhan and Schwing, 2018; Wang et al., 2019; Narasimhan et al., 2018; Zhu et al., 2020; Ziaeefard and Lécué, 2020; Yu et al., 2020; Li et al., 2020; Ramnath and and, 2020; Zhang et al., 2021a; Gardères et al., 2020; Zheng et al., 2021b; Vickers et al., 2021; Li and Moens, 2022; Zhang et al., 2022a; Ding et al., 2022; Hussain et al., 2022; Han et al., 2023; Song et al., 2023b; Ravi et al., 2023; You et al., 2023; Khademi et al., 2023; Yin et al., 2023a; Dong et al., 2024).
   - Identifying high-level attributes like scene names, object parts, and human activities, using various pre-trained classifiers or

APIs[9] (Wang et al., 2017; Wu et al., 2016; Narasimhan and Schwing, 2018; Wang et al., 2018a; Narasimhan et al., 2018; Yu et al., 2020; Ramnath and and, 2020; Marino et al., 2021; Hussain et al., 2022; Zhang et al., 2022a; Lin and Byrne, 2022; He and Wang, 2023; Song et al., 2023b; You et al., 2023; Khademi et al., 2023).

- Image captions and OCR text strings can be generated for information supplement (Wu et al., 2016; Su et al., 2018; Zhu et al., 2020; Yu et al., 2020; Salaberria et al., 2023; Chen et al., 2022c; Hussain et al., 2022; Gui et al., 2022; Lin and Byrne, 2022; Wu and Mooney, 2022; Lin et al., 2022; You et al., 2023; Zhou et al., 2023; Si et al., 2023; Khademi et al., 2023; Dong et al., 2024).
- Those question and captions can be parsed by NLP tools (e.g., NLTK (Bird et al., 2009), AllenNLP constituency parser (Gardner et al., 2018), Stanza (Qi et al., 2020), NLP Dependency Parser (Chen and Manning, 2014), Named Entity Recognizer (Finkel et al., 2005), LLMs (Dong et al., 2024)) for syntax analysis (Wang et al., 2019; Li et al., 2020; Saqur and Narasimhan, 2020; Cao et al., 2022b; Han et al., 2023; Wu and Mooney, 2022; Ravi et al., 2023), along with techniques like regular expressions (regex) (Wang et al., 2017), semantic graph parsing model (Zhu et al., 2020; Yu et al., 2020; Wu et al., 2022; He and Wang, 2023; Hussain et al., 2022), SpanSelector (Jain et al., 2021), or query template selector (Wang et al., 2018a; Narasimhan and Schwing, 2018; Narasimhan et al., 2018).
- Unimportant visual objects not present in the question or caption might be filtered out (Zhang et al., 2021a; Gardner et al., 2018).
- After extracting initial concepts from Q and I, two key mappings are established: the first links parsed objects in Q to their visual counterparts in I, and the second associates these concepts with relevant entries in KBs. This is achieved using methods like greedy longest-string matching (Wang et al., 2017; Su et al., 2018; Shevchenko et al., 2021), template matching (Wang et al., 2018a), and Multi-modal Entity Linking methods (Zheng et al., 2021b; Jain et al., 2021; Wu and Mooney, 2022; You et al., 2023; Adjali et al., 2023). Techniques like face identification algorithms (Shah et al., 2019; Vickers et al., 2021; Heo et al., 2022; Lerner et al., 2022) and ViLBERT-multi-task (Lu et al., 2020) serve as effective tools for linking objects.
- Fact triples can be collected by involving the first-order sub-KG from these identified concept nodes (sometimes will be three-hops in character KG (Shah et al., 2019)) or by identifying brief knowledge paths among the entities from I and Q. This process requires constructing a temporary (local) sub-KG specific to the current Q-I pair (Wang et al., 2019; Su et al., 2018; Li et al., 2020). In addition, KG-Aug (Li et al., 2020) constructs a global sub-KG that links Q, I, and candidate answers in a unified knowledge-based semantic space. KAN (Zhang et al., 2021a) presents a weighting system for each fact to indicate the reliability of the corresponding knowledge piece. Heo et al. (2022) develop a hypergraph from the KG, using a triplet as the basic unit to preserve the higher-order semantics inherent in the KG.
- RDF query (e.g., SPARQL) generation often involves filling pre-defined templates with parsed question data, suitable for datasets with consistent question patterns (Wang et al., 2017). Those queries typically include both "ASK" and "SELECT" types, with "ASK" checking for a solution to the query pattern and "SELECT" returning variables from all matched solutions (Wang et al., 2017).
- Term-based (e.g., TF-IDF and BM25) retrievers is another good choice, with their scoring reflecting the direct correlation between the query and fact triplets. Luo et al. (2021) use image captions generated by a model, concatenating them with Q as a query for BM25-based document retrieval. LaKo (Chen et al., 2022c) presents a Stem-based BM25 approach, using word stems as the smallest semantic units to maximize knowledge extraction from limited

[9] https://azure.microsoft.com/en-us/products/cognitive-services/vision-services

VQA and KG resources. EnFoRe (Wu and Mooney, 2022) utilizes entity-augmented queries to recall passages via BM25, measuring an entity's importance to the query by the relevance of these passages to the answer.

2. **Retrieval Pruning**:
   - Re-ranking candidate facts and may include assigning weights to nodes based on corresponding visual object sizes (Wang et al., 2019), ensuring each knowledge triple contains key elements from Q or auto-generated captions (Su et al., 2018; Wu and Mooney, 2022), or aligns with the relation type implied in Q (Narasimhan et al., 2018; Zhu et al., 2020; Yu et al., 2020; Ramnath and and, 2020; Hussain et al., 2022; Yin et al., 2023a).
   - A learnable score function can assess the compatibility between a fact representation and the Q-I representation (Narasimhan and Schwing, 2018; Hussain et al., 2022; Ravi et al., 2023).
   - Global KG-level pruning is also practiced. For example, KRISP (Marino et al., 2021) gathers all symbolic entities from the VQA dataset, including questions, answers, and visual concepts recognized by visual systems, and incorporates only triples related to these concepts to the model training. LaKo (Chen et al., 2022c) streamlines the KG by creating a stem corpus specific to the VQA field, ensuring all KG triples contain at least one stem from this corpus. KAT (Gui et al., 2022) extracts a subset from Wikidata (Vrandecic and Krötzsch, 2014) covering common real-world objects, and RR-VEL (You et al., 2023) only retains triples in the KG that include candidate answers and visually detected entities in the training set images.
3. **Search Engine**:
   - Marino et al. (2019) gather Wikipedia articles for each Q-I pair and select sentences closely matching the query based on key word frequency. Their ArticleNet predicts the presence and positioning of correct answers in these articles. Jain et al. (2021) utilize Google's search engine to retrieve the top-10 relevant snippets for a Machine Reading Comprehension (MRC) module based on a reformulated Q.
   - MAVEx (Wu et al., 2022) enriches knowledge retrieval through Google APIs for category labels, OCR readings, and logo information, collecting sentences from Wikipedia articles that contain candidate answers. It also uses Google Image Search with candidate Q-A pairs to provide additional visual information. Luo et al. (2021) notice that snippet-level knowledge outperforms sentence-level, and select ten snippets for each Q-A query.
4. **Dense Retrieval (Karpukhin et al., 2020)**:
   - This technique utilizes embedding similarities to match questions and visual concepts with pre-flattened concise fact sentences (Narasimhan et al., 2018; Zhu et al., 2020; Yu et al., 2020; Ziaeefard and Lécué, 2020; Wu et al., 2022; Li and Moens, 2022; Gao et al., 2022; Ossowski and Hu, 2023; Liu et al., 2022; You et al., 2023; Si et al., 2023).
   - Retrieval efficiency is frequently enhanced by employing open-source indexing engines like FAISS (Johnson et al., 2021), which facilitates the organization and indexing of large-scale dense embeddings. The architectures involved are generally symmetrical or siamese to support shared embedding spaces, while asymmetrical designs are adopted for Cross-Modal Retrieval scenarios (e.g., CLIP-based retrieval).
   - DMMGR (Li and Moens, 2022) ranks triplets based on the average cosine similarity between each word in the triplet and both the nouns in Q and the objects detected in I, excluding pairs with a zero average similarity.
   - RR-VEL (You et al., 2023) assesses the similarity between Q and key entities across various knowledge triples, using combined similarity scores to rank the candidate triples. KAT (Gui et al., 2022) uses the CLIP model to encode patch-level image regions and knowledge entries for retrieval purposes.
   - MAVEx (Wu et al., 2022) creates a concept pool for each Q-A pair, selecting facts containing potential answers identified by other VQA models. These facts are en-

coded using a pre-trained BERT model for re-ranking. Its subsequent work EnFoRe (Wu and Mooney, 2022) also prioritizes key entities in the Q-I pair, enhancing the knowledge retrieval process by focusing on entities crucial for answering the question.

- HKEML (Zheng et al., 2021a) applies 2D convolutional operations (Gao et al., 2021) to align the head and relation patterns in knowledge queries with Q, effectively mining implicit connections within the KG pertinent to Q-A pairs.

5. **Learnable Retriever**:
   - Learnable Retriever refers to a trainable retrieval model that enhances the adaptability and compatibility in KG-based VQA settings (Chen et al., 2021d; Luo et al., 2021; Lin and Byrne, 2022; Ravi et al., 2023; Wu et al., 2023b; Adjali et al., 2023).
   - Chen et al. (2021d) and Li et al. (2022b) separately aligning the joint embedding of the Q-I pair with the targets like relations in separate feature spaces. The prediction for relation type in the sub-KG pruning operation mentioned before is similar.
   - VLC-BERT (Ravi et al., 2023) assigns similarity scores to inference facts for each Q based on their overlap with human-annotated answers. These scores act as weak signals, indicating the relevance of each fact to Q, thus guiding the training of the retriever.
   - Luo et al. (2021) utilize DPR (Karpukhin et al., 2020) as a neural retriever, leveraging two BERT models for encoding the query and context. They further adapt DPR for visual domains with two variants: Image-DPR based on LXMERT (Tan and Bansal, 2019), and Caption-DPR, which modifies the DPR approach to suit visual content.
   - LaKo (Chen et al., 2022c) explores a differentiable KG retriever, leveraging cross-attention scores between the token of the prediction output and input facts for iterative reader-retriever training.
   - Addressing the challenge of slow convergence and sub-optimal performance in learnable retrievers, DEDR (Salemi et al., 2023a) employs a dual multi-modal encoder architecture with shared parameters for both Q-I queries and knowledge content, starting from the same shared embedding space. It further explores both multi-modal and text-only retrievers, combining their results via an ensemble method. The training of these retrievers utilizes a multi-modal retrieval dataset provided by Qu et al. (2021) as a supervised corpus. Training for these retrievers is based on a supervised multi-modal retrieval dataset from Qu et al. (2021).
   - REVEAL (Hu et al., 2023) integrates three data sources: WikiData KB (Vrandecic and Krötzsch, 2014), Wikipedia-Image-Text (WIT) (Srinivasan et al., 2021), and the VQA2.0 dataset (Antol et al., 2015). It utilizes a gating mechanism for optimal knowledge source selection and employs the perceiver architecture (Jaegle et al., 2021) to encode and compress knowledge items, enabling cascading multi-modal retrievers and joint reasoning.
   - RAVQA (Lin and Byrne, 2022) treats retrieval content as negative if it does not aid in answer generation, using the rest as positive samples. This approach helps in training the retriever by defining relevant and irrelevant content. Additionally, it combines the retrieval probability with the reader's answer prediction to determine the final result.
   - **Cold Start** issues: REVEAL (Hu et al., 2023) creates an initial retrieval dataset with pseudo ground-truth knowledge, using a large-scale image-caption dataset (Srinivasan et al., 2021). For pre-training, REVEAL pairs passages with query images as pseudo ground-truth knowledge and, to align with VQA task formats, randomly truncates captions to predict the truncated content using the image and the remaining text. LaKo (Chen et al., 2022c) initially employs a BM25-based retriever for knowledge retrieval in the first training phase, allowing for preliminary distillation to the differentiable retriever to mitigate the cold start problem.

6. **PLM Generation as the Retrieval**:
   - Implicit knowledge encoded in model parameters, typically pre-trained on large-scale datasets via self-supervised tasks,

makes the use of an explicit retriever optional for VQA completion. Several studies directly and implicitly harness the knowledge embedded in PLMs for reasoning, often skipping a separate knowledge retrieval step (Salaberria et al., 2023; Yang et al., 2022; Zhang et al., 2022a; Shao et al., 2023; Subramanian et al., 2023).

- KAT (Gui et al., 2022) and TwO (Si et al., 2023) take GPT-3 to retrieve implicit textual knowledge with supporting evidence; VLC-BERT (Ravi et al., 2023) uses COMET (Hwang et al., 2021), a LM trained on commonsense KGs, to generate contextual expansions instead of direct knowledge retrieval from KBs. Wang et al. (2023g) utilize ChatGPT to decompose Q, alleviating the issue of unfocused and lacking detailed image features in image captioning; MM-Reasoner (Khademi et al., 2023) employs LLMs to create rationales from multi-aspect visual descriptions (e.g., commonsense knowledge facts and external information). These rationales, alongside I and Q, are processed by a specifically fine-tuned Visual Language Model (VLM) designed to handle such enriched input.

Knowledge Representation involves selecting the appropriate format for symbolic KGs to integrate with multi-modal models. This decision is crucial for effectively infusing knowledge into multi-modal reasoning tasks.

1. **Direct Text-to-Embedding Mapping**:
   - Some research treats entities and relations in KGs as words, using embedding methods like Glove (Pennington et al., 2014) to translate them into continuous vectors. This transformation enables the further compression of knowledge components (e.g., triples) into fixed-size vectors using Recurrent Neural Networks (RNNs) (Wang et al., 2019), (V)PLMs (You et al., 2023; Hu et al., 2023; Chevalier et al., 2023), or mean pooling (Narasimhan and Schwing, 2018; Narasimhan et al., 2018; Zhu et al., 2020; Yu et al., 2020; Chen et al., 2021d; Marino et al., 2021; Li and Moens, 2022; Wu et al., 2022; Hussain et al., 2022).
   - When handling lengthy text from SPARQL queries, Wu et al. (2016) use Doc2Vec (Le and Mikolov, 2014) to learn feature representations for variable-length texts.
   - Techniques such as stop-word removal in Word2Vec can further refine knowledge representation, reducing the noise from irrelevant words in mean pooling (Narasimhan et al., 2018; Liu et al., 2022; Chen et al., 2022c).
   - Some methods convert fact collections into natural language sentences via concatenating the relation and object entities (Ziaeefard and Lécué, 2020; Zhang et al., 2021a; You et al., 2023; Hu et al., 2023), allowing direct encoding into fixed-length vectors by PLMs.
2. **Knowledge Graph Embedding (KGE)**:
   - KGE offers a practical approach to embed facts and reveal semantic relationships among triples in an abstract space. This technology is valuable for setting up initial (Su et al., 2018; Ramnath and and, 2020; Zheng et al., 2021b; Han et al., 2023) fact embeddings and gathering multi-modal knowledge (Ding et al., 2022).
   - Cao et al. (2022b) train the RotatE model on the entire KG to get entity and relation features, modifying a guided-attention block to fuse those knowledge embeddings with I and Q features.
   - Chen et al. (2021d) evaluate various embeddings including TransE-based KG embeddings, BERT-based node representations of ConceptNet (Yang et al., 2023b; Malaviya et al., 2020), and GloVe embeddings, finding that Word2Vec representations excel at mapping answers in smaller datasets.
   - RVL (Shevchenko et al., 2021) utilizes the PyTorchBigGraph method (Lerer et al., 2019) for embedding the Wikidata KG, while KVQAmeta (García-Olano et al., 2022) employs Wikipedia2Vec for representing entities from Wikipedia, emphasizing KGE's versatility in representing different knowledge sources.
3. **Pure Context**:
   - In many cases, KG triples are maintained in their original textual format for direct participation in multi-modal reasoning. They serialize triples for joint reasoning with (V)PLMs (Vickers et al., 2021; Chen et al., 2022c; Gao et al., 2022; Yang et al., 2022;

Gui et al., 2022; Lin and Byrne, 2022; Wu and Mooney, 2022; Lin et al., 2022; Si et al., 2023; Ravi et al., 2023; You et al., 2023; Shao et al., 2023; Zhou et al., 2023; Wang et al., 2023g; Hu et al., 2022; Xenos et al., 2023; Khademi et al., 2023; Dong et al., 2024).

Knowledge-aware Modality Interaction.

1. **Concatenation**:
   - This unified feature is typically refined with a Multi-Layer Perceptron (MLP) to enhance modality interaction and integration. In multi-modal fusion models like MUTAN (Ben-Younes et al., 2017), BAN (Kim et al., 2018), SAN (Yang et al., 2016) and ERMLP (Ramnath and and, 2020), feature concatenation is a preliminary step before being input to the MLP layer, crucial for sophisticated multi-modal analysis.
2. **Long Short-Term Memory (LSTM) Network**:
   - LSTM Network is a foundational framework for integrating knowledge with multi-modal data.
   - Sometimes LSTMs also act as standalone encoders for textual data (Narasimhan and Schwing, 2018; Narasimhan et al., 2018; Yu et al., 2020; Zhu et al., 2020; Li et al., 2020; Ramnath and and, 2020; Zhang et al., 2021a; Li and Moens, 2022), employing Glove (Pennington et al., 2014; Han et al., 2023) or PLMs (Devlin et al., 2019; Lan et al., 2020) for token embedding initialization. The output embeddings aid in subsequent stages of modality fusion, giving LSTM a pivotal role similar to those methods in *Direct Text-to-Embedding Mapping* paradigm.
3. **Graph Neural Networks (GNNs)**:
   - GNNs emphasize the connection of concepts in VQA by integrating representations from I, Q, and entities into cohesive networks, where each node (entity) is represented by an embedding that is a concatenation of different modalities (Narasimhan et al., 2018).
   - Mucko (Zhu et al., 2020) diverges from traditional modality embedding concatenation by independently processing distinct modalities' KGs. This involves isolating and separately analyzing the visual scene KG, the semantic KG from image captions, and the common sense KG, supporting precise answer determination through Q-guided attention and cross-KG convolution. The method of Q-guided KG node weighting has seen similar implementations in other studies (Yu et al., 2020; Li et al., 2020; Saqur and Narasimhan, 2020; Ziaeefard and Lécué, 2020; Li and Moens, 2022; Liu et al., 2022; Hussain et al., 2022; Wang et al., 2022b).
   - KG-Aug (Li et al., 2020) uses GCN to generate entity representations, which are then used to embed knowledge into the features of both Q and I.
   - KRISP (Marino et al., 2021) applies a RGCN (Schlichtkrull et al., 2018) for symbolic knowledge reasoning, enhancing each entity with four inputs: *a)* A binary indicator for concept presence in Q; *b)* Classifier probabilities for the concept's node, or zero if not detected in I, using various classifiers and detectors; *c)* A GloVe pooling representation of the concept; *d)* An implicit knowledge representation derived from a multi-modal pre-trained model (Li et al., 2019).
   - VQA-GNN (Wang et al., 2022b) employs a multi-modal GNN with bidirectional fusion to update concept and scene graph nodes for answer prediction through inter-modal message passing.
4. **Dynamic Memory Networks (DMNs)**:
   - DMNs (Kumar et al., 2016) utilize an attention-based mechanism for filtering critical information from localized small-scale knowledge triple embeddings (Fig. 4 (d)), achieved by modeling interactions across multiple data channels (Wang et al., 2019; Shah et al., 2019; Han et al., 2023; Yin et al., 2023a).
   - Through *triple replication*, VKMN (Su et al., 2018) deconstructs each knowledge triple into three Key-Value pairs, for instance, (h, r) as the key and t as the value, reducing interference caused by using only head and tail entities as keys for retrieval thereby improving reasoning performance.
   - DMMGR (Li and Moens, 2022) follows this setting and further refines knowledge triple composition by using the average em-

bedding of a triple as a key and its individual elements as values for enhanced relevance assessment. These networks apply a multi-scale attention mechanism that initially evaluates the overall relevance of a triplet's embedding, then assess the importance of each element, leading to more accurately recalled dynamic memories.

- GRUC (Yu et al., 2020) uses visual and semantic scene graphs as knowledge sources for external memory, iteratively updating multi-modal memories and employing a GRU module to refresh factual entity representations, incorporating inputs from previous entities and memory from the last time step.
- SUPER (Han et al., 2023) integrates a memory augmented component to retain and adjust key clues for answering questions, a method named *memory reactivation*. REVEAL (Hu et al., 2023) unifies multi-modal data by compressing each entry into a set number of value embeddings and a single key embedding for memory storage, achieving synchronous and stable updates between the memory encoder and main framework by re-encoding a portion (10%) of the retrieved knowledge items in each training iteration.

5. **Guided-Attention & Transformer**:
   - Many studies (Ramnath and and, 2020; Gardères et al., 2020; Zhang et al., 2021a; Cao et al., 2022b; Wu et al., 2022; Heo et al., 2022) have adopted a guided-attention mechanism to merge knowledge embeddings with visual and textual features.
6. **PLM & VLM Reasoning**:
   - **Embedding-Based Visual Information Integration**: This category includes methods that convert visual data into embeddings compatible with the input specifications of (V)PLMs (Dou et al., 2022). It involves techniques that restructure visual inputs into embeddings which seamlessly integrate with the model's existing architecture, such as compressing patch or local object features into fixed-length embedding sets (Jaegle et al., 2021; Hu et al., 2023) or applying adapters or projection heads for cross-modal feature space alignment (Lin et al., 2022; Yin et al., 2023b). These visual embeddings, combined with textual inputs, are processed in the embedding layers of (V)PLMs (Jaegle et al., 2021; Ossowski and Hu, 2023) as shown in Fig. 4 (f). Some studies (Vickers et al., 2021; Luo et al., 2021; Guo et al., 2022b; Ravi et al., 2023; Salemi et al., 2023a) integrate retrieved knowledge content and questions with image regions of interest, subsequently fine-tuning VLMs end-to-end on the VQA dataset using ground truth answers for optimization. RVL (Shevchenko et al., 2021) and KVQAmeta (García-Olano et al., 2022) inject the knowledge into the VLMs via aligning the KG embedding with the corresponding textual phrase representations derived from the output summations of PLM's embedding layers. MuKEA (Ding et al., 2022) uses the visual and language output sides of the LXMERT as the head and relation of a triple, respectively, pairing these with the ground truth answer as the tail entity. This association, aroused through the KGE method (e.g., TransE), leverages implicit knowledge within VLMs for reasoning. VLC-BERT (Ravi et al., 2023) uses a Q-guided multi-head attention block to fuse multiple knowledge representation vectors before feeding them into the VLM. He and Wang (2023) propose a graph-involved Q-attention mechanism, where V-Q guided graphs are built to direct VLM training by integrating a graph-aware mask matrix into the Transformers' attention matrix. Pang et al. (2023) enhance a VLM's ability for parametric knowledge injection by integrating the frozen Transformer layer of the LLM (LLaMA (Touvron et al., 2023)) between its cross-modal fusion and decoder modules.
   - **Textual Conversion of Visual Data**: This category involves converting all visual information into a textual format, like captions shown in Fig. 4 (f), enabling the application of PLM reasoning to a uniform textual dataset that includes background knowledge, questions, and images (Salaberria et al., 2023; Chen et al., 2022c; Luo et al., 2021; Zhang et al., 2022a; Yang et al.,

2022; Gao et al., 2022; Si et al., 2023; You et al., 2023; Zhou et al., 2023; Hu et al., 2022; Xenos et al., 2023). These works usually hold that text-only PLMs can effectively infer answers, even compensating for the loss of fine-grained visual features in image captions. Chen et al. (2022c) illustrate how encoder-decoder PLMs address the long-tail problem in answers and discrepancies between training and testing sets, while avoiding span prediction and directly generating free-form answers. Jain et al. (2021) reframe VQA as a MRC task, integrating search engines for additional context. TRiG (Gao et al., 2022) and TwO (Si et al., 2023) expand this approach to include object-level (e.g., object, attribute, and OCR labels) information alongside captions. Utilizing LLMs like GPT-3 with image captions, PICa (Yang et al., 2022) reveals that pure PLMs can achieve impressive performance in zero-shot and few-shot learning scenarios. KAT (Gui et al., 2022) further queries GPT-3 for providing reasoning evidence, aiming to extract deeper insights and implicit knowledge from GPT-3's outputs to bolster the reasoning process. REVIVE (Lin et al., 2022) employs a Transformer encoder as an adapter to utilize fine-grained regional visual information. PROOFREAD (Zhou et al., 2023) utilizes XGBoost (Chen and Guestrin, 2016), a gradient-boosted decision tree model, as a knowledge perceiver to classify knowledge entries based on their contribution scores across various dimensions. The Fusion-in-Decoder (FiD) approach (Izacard and Grave, 2021), where knowledge is individually compressed in the encoder and then jointly utilized in the decoder for reasoning, is adopted by various studies (Gui et al., 2022; Gao et al., 2022; Chen et al., 2022c; Wu and Mooney, 2022; Lin et al., 2022; Salemi et al., 2023a; Si et al., 2023). This allows for the simultaneous input of a large corpus of uni-modal or multi-modal background knowledge into the (V)PLMs.

- To mitigate the loss of fine-grained visual details in caption-based conversion, Wang et al. (2023g) leverage the LLM's reasoning capabilities to spotlight critical image details that that might be overlooked in captions. By decomposing the main Q into sub-questions and obtaining answers via a pre-trained VQA model, they identify and select those factual summaries with higher contribution scores than the original Q, supplementing the initial captions with these key details. This is similar to KAT (Gui et al., 2022) and TwO (Si et al., 2023), which apply In-Context Learning (ICL) in GPT-3 which employs a combination of Q, caption, and object labels as the prompt to generate implicit textual knowledge; PromptCap (Hu et al., 2022) introduces Q-guided caption generation to cover the visual details required by Q; ASB (Xenos et al., 2023) identifies the image patches most relevant to Q and generates informative captions from these patches only; Cola-FT (Chen et al., 2023d) prompts VLMs to generate captions and plausible answers separately, which are then concatenated with the instruction prompt, Q, and choices, forming a holistic context for LLMs to logically deduce the answer.

Knowledge-aware Answer Determination plays a crucial role in generating and predicting answers, often overlapping with *Knowledge-aware Modality Interaction*. Certain methods uniquely address both these aspects simultaneously, highlighting their intertwined nature.

1. **Information Extraction**:
   - To further rank the potential answers, some approaches implement heuristic rules, like matching score calculation (Wang et al., 2018a; Narasimhan and Schwing, 2018) and answer frequency assessment (Wang et al., 2018a).
2. **Discrimination**:
   - Such methods are effective when narrowing down potential answers within a certain range, often using GNN-alike models (Narasimhan et al., 2018; Liu et al., 2022; Hussain et al., 2022) as the backbones (Fig. 4 (c)).
   - Furthermore, discriminators can be either MLP-based (Wang et al., 2019; Narasimhan et al., 2018; Zhu et al., 2020; Yu et al., 2020; Liu et al., 2022) or rule-based (Narasimhan and Schwing, 2018). A

notable limitation of this approach is time consumption, especially when dealing with extensive answer vocabularies.

3. **Classification**:
   - Many studies reformulate the question-answering process as a classification problem, often employing a Fully Connected (FC) or MLP layer for answer prediction (Su et al., 2018; Marino et al., 2019; Shah et al., 2019; Li et al., 2020; Ziaeefard and Lécué, 2020; Saqur and Narasimhan, 2020; Zhang et al., 2021a; Gardères et al., 2020; Zheng et al., 2021b; Li and Moens, 2022; Guo et al., 2022b; Han et al., 2023; Heo et al., 2022; Li et al., 2022b; Wang et al., 2022b; Song et al., 2023b), where the output dimension corresponds to the pre-defined number of answer candidates.
   - Chen et al. (2021d) introduce an answer masking strategy that imposes direct knowledge-based constraints on the classifier's predicted answer probabilities, thereby limiting the range of potential answers. This method parallels KRISP (Marino et al., 2021), which employs late fusion to integrate the implicit and symbolic components of the model, selecting the highest-scoring answer from the combined answer vectors. MAVEx (Wu et al., 2022) introduces an answer validation module that leverages knowledge features from retrieved I, ConceptNet, and Wikipedia for answer candidate validation.
   - For (V)PLM-based methods, a classification (or projection) head is typically appended to the output `[CLS]` embedding (Fig. 4 (f)) (Shevchenko et al., 2021; Luo et al., 2021; Salaberria et al., 2023; García-Olano et al., 2022; Guo et al., 2022b; Ding et al., 2022; He and Wang, 2023; Ravi et al., 2023), often utilizing encoder-based backbones like LXMERT (Tan and Bansal, 2019) and BERT (Devlin et al., 2019). However, a significant trade-off common to classification-based approaches still exists, as noted by Chen et al. (2022c): the necessity to balance answer coverage and error rate, which hinges on pre-defining the answer candidate set according to its occurrence frequency.

4. **Generation**:
   - Textual Generative Models have become increasingly important in VQA tasks, particularly for addressing questions with answers outside pre-defined vocabularies. Generative (V)PLM-based methods are now increasingly supplanting traditional classification-based approaches (Chen et al., 2022c; Yang et al., 2022; Gao et al., 2022; Gui et al., 2022; Lin and Byrne, 2022; Wu and Mooney, 2022; Zhang et al., 2022a; Lin et al., 2022; Salemi et al., 2023a; You et al., 2023; Si et al., 2023; Hu et al., 2023; Shao et al., 2023; Zhou et al., 2023; Ossowski and Hu, 2023; Wang et al., 2023g; Hu et al., 2022; Xenos et al., 2023; Ghosal et al., 2023; Khademi et al., 2023).
   - These methods, using decoder-based or encoder-decoder based models like GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), VL-T5 (Cho et al., 2021), and BLIP-2 (Li et al., 2023b), feed constructed prompts to implicitly retrieve knowledge and perform analytical reasoning. Answer generation often relies on greedy decoding or beam search strategies (Gao et al., 2022; Ravi et al., 2023; Khandelwal et al., 2023), with the former selecting the most probable token at each step and the latter maintaining a fixed-size beam to produce a list of ranked answer candidates. To improve few-shot learning performance in models with large parameters, such as GPT-3, strategies like incorporating high-quality ICL examples (Shao et al., 2023; Wang et al., 2023g; Hu et al., 2022; Xenos et al., 2023) and employing multi-query ensembles (Yang et al., 2022; Xenos et al., 2023) are effective. Prophet (Shao et al., 2023) enhances this process by first generating candidate answers using a standard VQA model, subsequently refined through GPT-3. Meanwhile, Cola-FT (Chen et al., 2023d) prompts VLMs to generate captions and plausible answers separately, then integrating them with the instructional prompt, question, and candidate options for LLM-based reasoning.
   - As shown in Table 1, a noticeable increase in text-generation-based VQA methods is observed in the last two years. This trend

can also be attributed to the limitations of Exact Match answer evaluation manner in VQA benchmarks, which historically do not provide an advantage in evaluating the performance of open-ended generative models.

Recent advancements include CodeVQA (Subramanian et al., 2023), a training-free method prompting Codex (Chen et al., 2021c) with in-context examples to break down Q into Python code. This method leverages pre-defined visual modules in pre-trained VLMs, utilizing conditional logic and arithmetic. In line with NLP findings that LLMs improve performance at reasoning tasks when solving problems step-by-step (Wei et al., 2022; Yin et al., 2023b), VQA performance improves by decomposing Q and answering sub-questions sequentially. Khandelwal et al. (2023) propose *Successive Prompting*, where a LLM generates and resolves follow-up questions one at a time using a VLM, culminating in an answer to the original Q.

Metrics. In VQA performance evaluation, *Accuracy* (Acc), defined as the proportion of correctly answered test questions, is a predominant metric. The Georgia Tech Visual Intelligence Lab's VQA Python API[10] employs a standard technique for computing Acc is recommended in the VQA challenge (Antol et al., 2015):

$$\text{Acc}(ans) = \min(1, \#\{human\ that\ said\ that\ ans\}/3)\,. \quad (2)$$

This metric assigns a soft score (ranging from 0 to 1) to each answer, based on a voting mechanism among multiple annotators. In contrast, the Exact Match (EM) metric treats all annotated answers as ground truth (GT), offering a less stringent evaluation criterion (Gao et al., 2022). Additionally, the WuPalmer similarity (WUPS) (Wu and Palmer, 1994) calculates the similarity between words based on their common sub-sequences in a taxonomy tree. A candidate answer is considered correct if its similarity to a reference word exceeds a specified threshold. Chen et al. (2022c) introduce Inclusion-based and Stem-based Acc metrics. The former considers an answer A correct if it includes or is included by a GT answer after normalization. The latter assesses correctness based on the intersection of stems between A and the GT A (e.g., the stem of "*happy*" and "*happiness*" is "*happi*"). Not that other NLP automatic evaluation metrics, beyond assessing answer correctness, can also evaluate the model's explanation quality. For example, generative metrics such as BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and METEOR measure the linguistic quality and relevance of rationale statements against a reference set. Originally developed for machine translation, these metrics provide insights into the generated explanations' coherence and fluency, complementing the evaluation of answer correctness.

```
Question: What is he doing?
Answer: horseback riding
Candidate: riding a horse
Is the candidate correct? [yes/no]
```

Given the limitations of lexical matching metrics in evaluating open-domain VQA predictions from generative models, where entirely different words may convey the same meaning, (Khandelwal et al., 2023), Kamalloo et al. (2023) further propose an evaluation metric that leverages InstructGPT (Ouyang et al., 2022), prompting it with Q and the candidate answers to determine their correctness. An example of this process is illustrated in the adjacent code snippet.

Knowledge Base. The background KBs for knowledge-aware multi-modal reasoning frequently involves multiple KGs, each bringing its unique and complementary insights to the reasoning process. **Trivia** knowledge, such as DBpedia, provides facts about famous people, places, and events. **Commonsense** knowledge, offers insights into basic concepts like the composition of houses or parts of a wheel. **Scientific** knowledge, found in databases like hasPart KB, details classifications and properties, such as the genus of dogs or types of nutrients. Lastly, **situational** knowledge from resources like Visual Genome offers contextual data, e.g., typical locations of cars or common contents found in bowls.

- **ConceptNet** (Speer et al., 2017) encapsulates human commonsense knowledge, containing various relations including *usedFor*, *createdBy*, and *isA*, primarily generated from the Open Mind Common Sense (OMCS) project;
- **DBpedia** (Auer et al., 2007), constructed from Wikipedia, spans multiple fields relevant to daily life. In this KG, concepts are connected through categories and super-categories in accordance with the SKOS[11] Vocabulary;

[10] https://github.com/GT-Vision-Lab/VQA

[11] http://www.w3.org/2004/02/skos/

- **WebChild** (Tandon et al., 2014) connects nouns with adjectives through fine-grained relations, such as *hasShape*, *faster*, *bigger*. This information is automatically extracted from the Web;
- **Wikidata** (Vrandecic and Krötzsch, 2014) offers extensive factual knowledge, including a broad range of topics about the world;
- **hasPart KB** (Bhakthavatsalam et al., 2020) documents relationships between objects, both common and scientific, such as *(Dog, hasPart, Whiskers)* and *(Molecules, hasPart, Atoms)*;
- **Visual Genome (VG)** (Krishna et al., 2017) gathers scene graphs from real-life situations, focusing on spatial relationships, e.g., *(Boat, isOn, Water)*, and common affordances, e.g., *(Person, sitsOn, Couch)*;
- **ATOMIC** (Hwang et al., 2021) consists of over 1M knowledge triplets covering a range of topics, including physical-entity relations, event-centered relations, and social interactions.
- **CSKG** (Ilievski et al., 2021) is a large consolidated source that integrates commonsense knowledge from seven diverse and disjoint sources, including ConceptNet, Wikidata, ATOMIC, VG, Wordnet (Miller, 1995), Roget (Kipfer, 1992) and FrameNet (Baker et al., 1998).

BENCHMARKS: We select FVQA (Wang et al., 2018a) and OKVQA (Marino et al., 2019) as our primary datasets due to their critical contributions to advancing knowledge-aware VQA and their significant impact on the development of subsequent datasets. Table 1 presents a chronological analysis of relevant methods, detailing their performance, model paradigms, and design principles.

Discussion 1 *VQA datasets vary in their answer formats, ranging from multiple-choice, where models select from provided options, to open-ended formats that test a model's understanding, reasoning, and independent answer generation or retrieval capabilities. Beyond answer formats, an important consideration in these datasets is the use of a Ground Truth (GT) set of facts for answering questions. Datasets like those in the FVQA series come with their own GT facts, while those in the OKVQA series do not. These facts should ideally be employed not for training purposes (such as pre-training a relation classifier) but for assessing the model's proficiency in KG fact retrieval. Besides, selecting appropriate knowledge sources and methods for knowledge filtering is also crucial for model performance.*

Furthermore, as indicated in Table 1, the comparison of VQA works can be influenced due to varying background KG sources and backbone models. Ensuring consistency in these aspects is essential for fair comparative analysis. It's important for researchers to distinguish whether improvements are due to the quality of the KB, the method of KG integration, or the backbone's inherent capabilities. These distinctions, often overlooked, are crucial to understanding genuine progress in the field. Relying solely on sophisticated visual, language, or multi-modal backbones to claim SOTA results, without addressing the uniformity of model parameters and ensuring fair comparisons, may compromise the credibility of the findings. Given VQA's practical applications, additional factors such as time, space complexity, real-time consumption, and GPU requirements are also significant for a comprehensive evaluation of these models.

Resource: In analyzing the evolution of KG-aware VQA datasets, we categorize the developments into three main groups: FVQA-type, OKVQA-type, and others.

***(i)* FVQA (Wang et al., 2018a):** The **KB-VQA** dataset (Wang et al., 2017) first evaluates VQA algorithms' ability to leverage external knowledge for answering complex image-based questions. It consists of multiple Q-A pairs per image, crafted by five questioners using predefined templates. These pairs aim to probe knowledge levels that surpass mere visual observation by leveraging DBpedia as the knowledge source. Expanding on KB-VQA, FVQA (Wang et al., 2018a) includes more questions, images and integrates additional KGs such as ConceptNet and Webchild. Notably, FVQA is the first VQA dataset to provide supporting facts for each question (i.e., external knowledge facts, rather than visual relation facts in R-VQA (Lu et al., 2018b)), paving the way for developing more knowledgeable VQA systems. **Variants: ZS-F-VQA** (Chen et al., 2021d) targets Zero-shot VQA, designed to prevent overlap between training and testing answers, paying attention on answer bias and Out Of Vocabulary (OOV) issues. **KRVQA** (Cao et al., 2022b) imposes constraints to promote image context engagement over mere knowledge fact memorization; **FVQA 2.0** (Lin et al., 2023) increases dataset size and introduces adversarial question variants to balance the answer

Table 1: Comparison of Knowledge-based VQA accuracy results on OKVQA (Marino et al., 2019) and FVQA (Wang et al., 2018a). The icon ◐ represents KG-based VQA methods; ◑ indicates methods without KG utilization. The † symbol signifies methods pre-trained on VQA2.0 or similar datasets. ⋆ indicates results reported on dataset version 1.1, which differs from version 1.0 in answer stemming methods. Abbreviations used: Q (Question), V (Visual), w/ (with), KG (Knowledge Graph), CN (ConceptNet), WP (Wikipedia), WC (WebChild), WD (Wikidata), DBP (DBpedia), VG (VisualGenome), YG (YAGO), HP (hasPart KB), AT (ATOMIC (Sap et al., 2019)), AS (Ascent (Nguyen et al., 2021)), VLM (Visual-Language Model), GNN (Graph Neural Network), GAT (Graph Attention Network), MRC (Machine Reading Comprehension), MHA (Multi-head Attention), DMN (Dynamic Memory Network), DPR (Dense Passage Retriever), FiD (Fusion-in-Decoder), In-context Learning (ICL), GI (Google Image), GS (Google Search), Enc.Dec.(Encoder-Decoder), DC (Discrimination), IE (Information Extraction), CLS (Classification), TG (Text Generation), WIT (Wikipedia-Image-Text (Srinivasan et al., 2021)). For methods employing both PLM intrinsic knowledge and external KB, only the KB is listed in the knowledge source.

	Methods	Approaches (Paradigms)	Key Idea	Knowledge Source	FVQA	OKVQA
2018~2021	◐ QQmaping (Wang et al., 2018a)	RDF Query (IE)	Question-Query Mapping	DBP / CN / WC	56.91	-
	◐ STTF (Narasimhan and Schwing, 2018)	Relation Query (IE)	Scoring the Facts	DBP / CN / WC	62.20	-
	◐ OB (Narasimhan et al., 2018)	Fact Retrieval + GNN (DC)	Entity Graph + GCN	DBP / CN / WC	69.35	-
	◑ Marino et al. (2019)	Retrieval + ArticleNet (IE)	Web Retrieval + Knowledge Span Prediction	WP	-	27.84
	◐ KG-Aug (Li et al., 2020)	Fact Retrieval + GNN (CLS)	Augment Q&V Features w/ Entity Embedding	WD / CN	38.58	26.71
	◐ Chen et al. (2021d)	Alignment + Re-rank (CLS)	KG-aware Answer Masking for Validation	DBP / CN / WC	58.27	-
	◐ ERMLP (Ramnath and and, 2020)	KGE + Attention (IE)	Knowledge-guided Co-attention	DBP / CN / WC	60.82	-
	◐ Liu et al. (2022)	Fact Retrieval + GNN (DC)	Q-V Guided Cross-modal GAT	DBP / CN / WC	63.56	29.43
	◐ KAN (Zhang et al., 2021a)	Retrieval + Attention (CLS)	Question-guided MHA	CN	66.39	-
	◐ Mucko (Zhu et al., 2020)	Fact Retrieval + GNN (DC)	Question-guided Attention + Cross-KG GAT	DBP / CN / WC	73.06	29.20
	◐ GRUC (Yu et al., 2020)	Fact Retrieval + DMN (DC)	Fact-centered DMN + GRU	DBP / CN / WC	79.63	29.87
	◐ ConceptBert (Gardères et al., 2020)	GCN + Attention (CLS)	Compact Trilinear Interaction + MHA	CN	-	33.66
	◐ KRISP† (Marino et al., 2021)	GCN + VLM (CLS)	Global KG + RGCN + VisualBERT	HP / DBP / CN / VG	-	38.90⋆
	◐ MAVEx† (Wu et al., 2022)	Multi-retrieval + Re-rank (CLS)	Web Retrieval + VilBERT + Answer Validation	WP / CN / GI	-	38.70⋆
	◐ RVL† (Shevchenko et al., 2021)	KGE Alignment + VLM (CLS)	Aligning VLM Text Embedding w/ KGE	WD / CN / PLM	54.27	39.04
	◑ Luo et al. (2021)†	Dense Retriever + PLM (IE)	RoBERTa + DPR Learning + MRC	GS	-	39.20⋆
	◐ PGVQA (Song et al., 2023b)	Retrieval + Re-rank (CLS)	KG-aware Answer Refinement	DBP / CN / WC	-	41.07
2022	◐ SUPER (Han et al., 2023)	Multi-modules + DMN (CLS)	Q-V Guided Modular Routing + DMN	CN	48.90	30.46
	◐ MKRE (Hussain et al., 2022)	Fact Retrieval + GNN (DC)	Question Guided Attention + Cross-KG GAT	DBP / CN / WC	73.06	-
	◐ DMMGR (Li and Moens, 2022)	Retrieval + DMN + GNN (CLS)	Caption + Multi-scale DMN + GAT	DBP / CN / WC	81.20	-
	◑ CBM-BERT† (Salaberria et al., 2023)	Caption + PLM (CLS)	Caption + BERT + Ensemble	PLM	-	36.00⋆
	◑ CBM-T5† (Salaberria et al., 2023)	Caption + PLM (TG)	Caption + T5 + Ensemble	PLM	-	40.80⋆
	◐ UnifER† (Guo et al., 2022b)	Fact Retrieval + VLM (CLS)	Loss Gap Driven DPR Learning + ViLT	CN	-	42.13⋆
	◑ MuKEA† (Ding et al., 2022)	VLM + KGE	LXMERT + Multi-modal TransE	VLM	-	42.59⋆
	◑ PICa† (Yang et al., 2022)	Caption + Decoder (TG)	Caption + Tag + ICL + GPT3	GPT3	-	43.30⋆
	◐ KVQAmeta† (García-Olano et al., 2022)	KGE Alignment + VLM (CLS)	Aligning VLM Embedding w/ Wikipedia2Vec	WP	-	43.67⋆
	◐ LaKo† (Chen et al., 2022c)	Retrieval + Enc.Dec. (TG)	Caption + Fact DPR + T5 + FiD	HP / DBP / CN / WC	-	47.01⋆
	◑ TRiG† (Gao et al., 2022)	Retrieval + Enc.Dec. (TG)	Caption + Tag + DPR + FiD	WP	-	49.24⋆
	◐ KAT† (Gui et al., 2022)	Retrieval + Enc.Dec. (TG)	Caption + Tag + ICL + GPT3 + DPR + FiD	WD / GPT3	-	53.09 ⋆
	◐ EnFoRe† (Wu and Mooney, 2022)	Retrieval + Enc.Dec. (TG)	KAT + Entity Focused DPR	WD / GPT3	-	54.35⋆
	◑ RAVQA† (Lin and Byrne, 2022)	Retrieval + Enc.Dec. (TG)	DPR Learning + Ensemble	GS	-	54.48⋆
	◐ REVIVE† (Lin et al., 2022)	Retrieval + Enc.Dec. (TG)	KAT + Regional Visual	WD / GPT3	-	56.604⋆
2023	◐ VLC-BERT† (Ravi et al., 2023)	Fact Generation + VLM (CLS)	COMET Generation + MHA + VL-BERT	CN / AT	-	43.14⋆
	◑ DEDR† (Salemi et al., 2023a)	Retrieval + Enc.Dec. (TG)	Mutual Retriever Distillation + VL-T5 + FiD	WP	61.80	44.57⋆
	◐ RR-VEL† (You et al., 2023)	EL + Retrieval + Enc.Dec. (TG)	Ground Truth Referent in Q + T5	HP / CN / Ascent	65.59	49.48⋆
	◐ MCR-MemNN (Yin et al., 2023a)	Fact Retrieval + DMN (CLS)	Multi-clue Reasoning + Fact-centered DMN	DBP / CN / WC	70.92	-
	◑ CodeVQA† (Subramanian et al., 2023)	Code Generation + PLM (TG)	ICL + Codex + Modular Combination	Codex / VLM	-	53.50⋆
	◑ TwO† (Si et al., 2023)	Retrieval + Enc.Dec. (TG)	KAT + OFA Multi-modal Knowledge	WP / GPT3	-	57.57⋆
	◑ PROOFREAD† (Zhou et al., 2023)	Decoder + Re-rank (TG)	Answer-aware Knowledge Generation & Filter	ChatGPT	-	57.60⋆
	◐ REVEAL† (Hu et al., 2023)	Retrieval + Enc.Dec. (TG)	Multi-modal Retrieval + Gate + DMN + T5	WIT / WD / VQA2.0	-	59.10⋆
	◐ MM-Reasoner† (Khademi et al., 2023)	Fact Generation + Enc.Dec. (TG)	Vision APIs + ICL Rationales Generation	GPT-4	61.10	59.20⋆
	◑ Wang et al. (2023g)†	Q-decomposition + Decoder (TG)	Q Decomposition + Fact Refinement + PICa	ChatGPT	-	59.34⋆
	◑ PromptCap† (Hu et al., 2022)	Caption + Decoder (TG)	Q-guided Caption Generation + PICa	GPT-3	-	60.40⋆
	◑ Prophet† (Shao et al., 2023)	Caption + Decoder (TG)	MCAN + Answer Pruning + Answer-aware ICL	GPT3	-	61.10⋆
	◑ ASB† (Xenos et al., 2023)	Caption + Decoder (TG)	Q-guided Patch Caption Selection + PICa	LLaMA-13B	-	61.20⋆
	◑ Cola-FT† (Chen et al., 2023d)	Decoder (TG)	OFA + BLIP + LLM Answer Decision	FLAN-T5	-	62.40⋆
	◑ GPT-4V† (Li et al., 2023f)	Decoder (TG)	Prompt + GPT-4V	GPT-4V	-	64.28⋆

distribution;

***(ii)* OKVQA (Marino et al., 2019):** Different from FVQA, OKVQA dataset focuses on open-world VQA, involving questions that implicitly require external knowledge without specifying a direct KB link or providing explicit KG triplets. Its broad knowledge scope makes it a benchmark alongside the VQA2.0 dataset (Antol et al., 2015). **Variants:** **OKVQA**$_{S3}$ and **S3VQA** (Jain et al., 2021) enhance the original OK-VQA by incorporating questions that require object detection within images, with subsequent substitution of the detected object in the query and employing web searches to find answers; **A-OKVQA** (Schwenk et al., 2022) introduces a greater diversity of world knowledge and more reasoning steps to extend OK-VQA, further providing rationales for each question to aid in training explainable VQA models. **OKVQA2.0** (Reichman et al., 2023) refines OK-VQA with corrections and attaching Wikipedia sources to Q-I pairs; **ConceptVQA** (Gan et al., 2023) enriches OK-VQA with entity-level annotations aligned with ConceptNet entities and presents a unique challenge by ensuring its testing split features non-overlapping answers with the training set, similar to ZS-F-VQA (Chen et al., 2021d).

(iii) Others: Li et al. (2017) develop **Visual7W+KB** from the Visual7W test split images (Zhu et al., 2016), automating question creation using predefined templates and ConceptNet (Speer et al., 2017) for guidance; **KVQA** (Shah et al., 2019) incorporates world knowledge about named entities like *Barack Obama* and the *White House* from Wikidata (Vrandecic and Krötzsch, 2014), also employing face identification technology in image analysis; **ViQuAE** (Lerner et al., 2022) extends KVQA's scope to include a broader range of entity types beyond just persons; **VCR** (Zellers et al., 2019) targets understanding human intentions in movie scenes with questions such as "*why is* `[PERSON]` *doing this?*"; **AI-VQA** (Li et al., 2022b) utilizes Visual Genome scene graphs and ATOMIC KG (Hwang et al., 2021) event knowledge, enriched by including volunteer-annotated QA pairs and detailed scene/object descriptions; **DANCE** (Ye et al., 2023) re-formats knowledge triples as natural language riddles paired with images, aiming to infuse visual language models with commonsense knowledge; Gao et al. (2023) introduce **LoRA**, a dataset focusing on formal and complex description logic reasoning in VQA. Centered around a KB related to food and kitchen scenarios, LoRA aims to enhance the logical reasoning capabilities of VQA models, which are not adequately assessed by existing VQA datasets; **ScienceQA** (Lu et al., 2022), sourced from elementary and high school science curricula, includes $21,208$ items along with lectures and explanations. It challenges models to generate coherent explanations across a wide range of subjects, setting it apart from OKVQA. Despite not incorporating a KG in its design, ScienceQA is pivotal for advancing knowledge-intensive multimodal models, which marks a significant step in the evolution of future KG-aware VQA methods.

In addition, KG-aware VQA can also extend to various scenarios beyond traditional settings. For example, **KnowIT** VQA (Garcia et al., 2020) contains video clips from "*The Big Bang Theory*" with associated knowledge-based QA pairs, annotated by those dedicated fans well-versed in the show's content. **K-EQA** (Tan et al., 2023) employs a KB and 3D scene graphs, enabling an AI agent to navigate environments and answer environment-aware natural language queries.

A.3.2 Visual Question Generation

VQG (Xie et al., 2022; Chen et al., 2023b; Salemi et al., 2023b) leverages visual cues to generate questions, diverging from traditional VQA by prioritizing question creation. This process is crucial in educational applications, such as engaging children with questions about images to support learning. Early VQG models (Mostafazadeh et al., 2016) utilize RNNs to generate questions based solely on images, leading to questions that often lack specific focus. In the KG-aware VQG domain, volunteers create the K-VQG dataset (Uehara and Harada, 2023) by integrating external knowledge from resources like ConceptNet and Atomic (Hwang et al., 2021) with image content, using partially masked commonsense triplets to enrich questions with knowledge. Xie et al. (2022) develop a pipeline comprising a visual concept feature extractor, knowledge representation extractor, target object extractor, and a decoder. This setup, aligned with the process outlined in Fig. 3, integrates non-visual knowledge into VQG and employs FVQA for its evaluation. KECVQG (Chen et al., 2023b) utilizes a causal graph to analyze and correct spurious correlations in VQG by linking unbiased features with external knowledge, thereby disentangling visual features to lessen the impact of these correlations. Unlike VQA, VQG methods prioritize evaluation on mean-

ingfulness, logical soundness, and consistency with target knowledge over strict correctness (Uehara and Harada, 2023), often using NLP-style metrics like BLEU and CIDEr for assessment.

Discussion 2 *Evolving intelligent dialogues with chatbots remains a critical objective in VQG, particularly in empowering robots to formulate precise, knowledge-enriched questions that boost problem-solving capabilities for future advancements. Equally important is the progression towards more interactive KG-aware VQG systems, which can dynamically adapt their questioning strategies based on user interactions and feedback, marking a significant direction for future research. Moreover, with the ongoing rapid developments in VQA, transferring and adapting common problems and methods from VQA to VQG can catalyze further innovative breakthroughs in question generation technologies.*

A.3.3 Visual Dialog

VD (Chen et al., 2022a) extends the VQA task by adopting a multi-round format where a continuous series of Q-A pairs revolves around a single image. This setting shifts from the single-question focus of VQA to a dynamic, conversational interaction about the image, posing a challenge for agents to adaptively interpret evolving relationships among visual elements based on the dialogue context. VD methods typically leverage historical dialogue information as background knowledge (Guo et al., 2020; Jiang et al., 2020; Kang et al., 2021; Wang et al., 2022c; Zhao et al., 2023), employing visual graph construction, query-guided relation selection and GNN propagation for dialog reasoning. Guo et al. (2020, 2022a) introduce Q-conditioned attention to aggregate textual context from dialogue history, constructing a context-aware object graphs for Q-guided message passing. Similarly, KBGN (Jiang et al., 2020) uses cross-modal GNNs to bridge modal gaps and capture inter-modal semantics, retrieving information relevant to the current question from both vision and text sources.

Addressing the limitations of relying solely on internal knowledge from images and dialog history, some approaches integrate commonsense knowledge for enhanced conversational depth. These methods all align with the *KG-aware Understanding and Reasoning* paradigms we have previously outlined (Fig. 3). For example, ***(i)*** **Knowledge Retrieval**: SKANet (Zhao et al., 2021a) integrates commonsense knowledge from ConceptNet into VD by using concept recognition and n-gram matching techniques to build a sub-KG. ***(ii)*** **Knowledge Representation**: KACI-Net (Zhang et al., 2023e) selects triplets with at least two entities or relations mentioned in a questionand transforms them into textual format for subsequent processing. ***(iii)*** **Knowledge-aware Modality Interaction**: RMK (Zhang et al., 2022b) utilizes caption-based dense retrieval to fetch relevant facts from ConceptNet, injecting knowledge into the dialogues through sentence-level and graph-level cross-modal attention and embedding concatenation. ***(iv)*** **Knowledge-aware Answer Determination**: Acknowledging the issue of spurious correlations from unobserved confounders in retrieved knowledge, Liu et al. (2023a) construct a counterfactual commonsense-aware VD causal graph. This graph applies counterfactual reasoning to mitigate commonsense bias, reducing the effect of misleading or inaccurate commonsense in answer derivation.

Discussion 3 *Currently, the emphasis in knowledge-based VD mainly lies in using external commonsense knowledge, while other knowledge types, such as scientific and situational, have been relatively underexplored. However, the rise of LLMs is diminishing the distinction between VD and VQA, with In-context Learning techniques in VQA starting to overshadow the traditional role of context in dialogues. This shift prompts a need to reassess VD's unique contribution and its path forward. As the boundaries between VD and VQA continue to merge, identifying and articulating the distinct potential of VD becomes imperative.*

A.4 Supplement for KG-driven Multi-modal Classification Tasks

Discussions are also extended to related multi-modal tasks like Fake News Detection and Movie Genre Classification, highlighting the diversity and wide-ranging applications within the field.

A.4.1 Supplementary Information for IMGC

Image Classification (IMGC) aims to identify objects within images and, with deep learning advancements, has even surpassed human performance in challenges like ImageNet ILSVRC (Russakovsky et al., 2015). Consider a set of labeled training samples $\mathcal{D}_{tr} = \{(x, y)|x \in \mathcal{X}, y \in \mathcal{Y}\}$, a classifier aims to approximate a function $f : x \rightarrow y$ from input x to output label y with the assist of

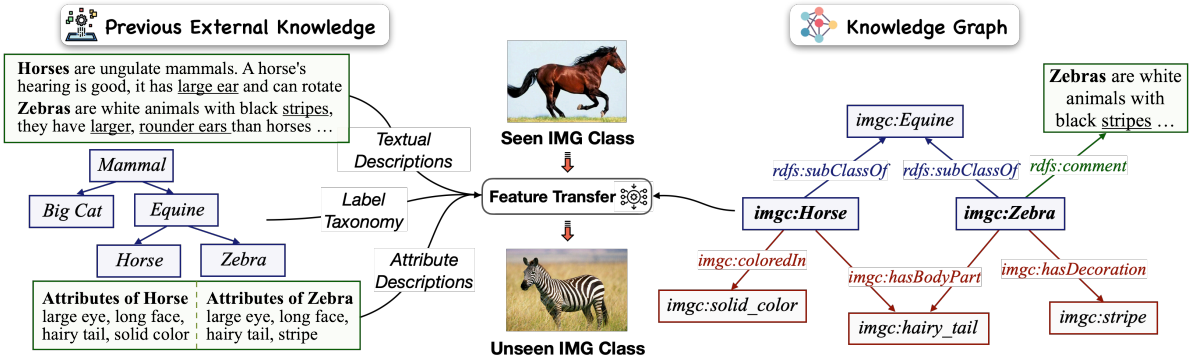


Figure 10: Comparison of previously used external knowledge (Left) and KG (Right) in Zero-shot Image Classification task (Geng et al., 2021a).

a background KG $\mathcal{G}$. This function should accurately predict the labels of samples in a testing set $\mathcal{D}_{te} = \{(x, y)|x \in \mathcal{X}', y \in \mathcal{Y}\}$, with $\mathcal{X} \cap \mathcal{X}' = \emptyset$.

In IMGC, x denotes an image, and y represents its class. Traditional IMGC follows a closed-world assumption, requiring extensive labeled images for both training and testing within known classes, i.e., $\mathcal{Y}_{tr} = \mathcal{Y}_{te} = \mathcal{Y}$. However, it is not feasible for newly emerging classes due to the impracticality of continuously annotating and retraining models with sufficient images for these classes. Consequently, there is a growing interest in Zero-Shot Image Classification (ZS-IMGC), which supports classifying images of novel, unseen classes without the need for specific training images, i.e., $\mathcal{Y}_{tr} \cap \mathcal{Y}_{te} = \emptyset$.

To handle these unseen classes, most existing ZS-IMGC methods adopt a knowledge transfer strategy (Chen et al., 2023c, 2021b): transferring labeled images, image features or model parameters from the seen classes in training set to unseen classes, guided by external knowledge that describes semantic relationships between classes. For instance, as illustrated in the left part of Fig. 10, consider the description of a "*Zebra*" as an animal with a horse-like body, tiger-like stripes, and black-and-white colors similar to a panda. Models can infer the appearance of a "*Zebra*" by combining features of these seen animals, even without direct exposure to its images. Briefly, ZS-IMGC relies on data from observed classes and class-specific semantic knowledge, with the external knowledge frequently embodying a modality distinct from the image data. This section reviews KG-based ZS-IMGC efforts to illustrate the typical practice of multi-modal learning in IMGC.

In ZS-IMGC literature, various forms of external knowledge are employed. Early ZS-IMGC works (Frome et al., 2013; Zhu et al., 2018) use textual class descriptions or names to model inter-class relationships. Others (Xian et al., 2018; Lampert et al., 2014) utilize class attributes, annotating each class with descriptive characteristics, thereby defining semantic relationships through shared attributes (see Fig. 10, left). However, these approaches sometimes face limitations in capturing complete semantics (Geng et al., 2023).

In KG-aware ZS-IMGC, a KG is defined as $\mathcal{G} = \mathcal{E}, \mathcal{R}, \mathcal{T}$, with $\mathcal{Y} \subset \mathcal{E}$. This paradigm, representing semantic hierarchical relationships among classes, is instrumental in augmenting classification performance and interpretability. For example, studies like (Wang et al., 2018b; Kampffmeyer et al., 2019) integrate hierarchical relationships from WordNet, while (Roy et al., 2022; Nayak and Bach, 2022; Gao et al., 2019) explore class knowledge from commonsense KGs such as ConceptNet. KGs, due to their compatibility, can unify various knowledge forms, including textual description and discrete attributes, into a single graph (see Fig. 10, right).

Mapping-based Methods. Chen et al. (2020a) employ an OWL-based ontology for animal classes, encoding it via the OWL EL embedding method and learn a linear encoder to map image features to class embeddings, with a focus on reconstruction loss for reverse mapping. Kata et al. (Zeynep Akata and Florent Perronnin and Zaïd Harchaoui and Cordelia Schmid, 2016; Akata et al., 2013) map initial class encodings into the image feature space, representing classes as multi-hot vectors of ancestors based on class hierarchies. There are also some joint mapping methods that map both the class encoding and the image features. For example, DUET (Chen et al., 2023h), an end-to-end Transformer-based ZSL method, leverages cross-modal PLMs for fine-grained visual characteristic reorganization and discrimination with structured KGs serialized as input.

Although mapping-based methods, which often employ linear or nonlinear transformations, are straightforward to implement, they exhibit a bias towards seen classes when trained exclusively on their images. This bias is particularly problematic in generalized ZSL scenarios, where seen and unseen classes coexist, highlighting their inherent limitations.

Data augmentation Methods. DOZSL (Geng et al., 2022) employs a disentangled KG embedding module to enhance the quality of synthesized image features in OntoZSL (Geng et al., 2021a). TGG (Zhang et al., 2019) generates few-shot samples, where a GAN-based generation module is applied to generate an image-level graph. Zhang et al. (2022d) develop a MMKG by merging visual representations of classes with word embeddings, creating multi-modal class nodes and edges that explicitly model class correlations, thereby enhancing information transfer from seen to unseen class nodes.

Propagation-based Methods. Certain methods tackle KGs' diverse relation types by using multi-relational GCNs (Chen et al., 2020b) or dividing multi-relation KGs into single-relation graphs with parameter-shared GCNs for feature propagation (Geng et al., 2022; Wang and Jiang, 2021; Wu et al., 2023a). For multi-label images, where models assign a probability score per class, GNNs utilize KG-implied correlations to propagate these scores from seen to unseen class nodes, as done by Lee at al. (Lee et al., 2018).

Discussion 4 *Incorporating diverse class semantics generally yield better ZS-IMGC results, even with basic methods. For instance, Table 2 illustrates that some mapping-based methods (Frome et al., 2013; Chen et al., 2020a), employing straightforward linear mapping functions, achieve better performance by integrating various class semantics such as attributes, hierarchy, and names, compared to GCNZ†, which relies solely on class hierarchy. Significantly, enriching GCNZ† with more class semantics, markedly enhances its effectiveness (GCNZ‡). Furthermore, KG-based ZS-IMGC methods typically operate in a class transductive setting, where unseen* ***classes*** *are known during training (Fig. 5 (b)), in contrast to the conventional inductive approach that utilizes only seen class knowledge. These methods leverage a KG to bridge seen and unseen classes through semantic links. Additionally, although the generalized ZSL setting is well-recognized by many researchers, some studies adopt their own definitions, specifically testing only unseen class images while classifying them within a combined pool of both seen and unseen classes. This variation requires careful consideration in future work.*

Table 2: Comparison of ZS-IMGC results across various datasets. We use Acc and H as the evaluation metrics of Standard ZSL and Generalized ZSL, respectively. DeViSE here is implemented with a KG which covers the semantics of class hierarchy, class attributes, attribute hierarchy and class names (see (Geng et al., 2023) for details). GCNZ† utilizes a KG with class hierarchy only, whereas GCNZ‡ includes broader class semantics like attributes, as reported in (Geng et al., 2023). DOZSL (GAN) and DOZSL (GCN) are variants of DOZSL (Geng et al., 2022), representing generation-based and propagation-based ZSL learners, respectively. ImageNet results are tested on 2-hops unseen classes.

Dataset	Methods	Acc (%)	H (%)
ImageNet	GCNZ (Wang et al., 2018b)	19.8	–
	DGP (Kampffmeyer et al., 2019)	26.6	–
	FGP (Wu et al., 2023a)	26.4	–
ImNet-A	DeViSE (Frome et al., 2013)	33.62	26.01
	OntoZSL (Geng et al., 2021a)	39.00	32.15
	DOZSL (GAN) (Geng et al., 2022)	40.26	32.82
	DOZSL (GCN) (Geng et al., 2022)	38.69	32.12
	GCNZ† (Wang et al., 2018b; Geng et al., 2023)	33.95	26.68
	GCNZ‡ (Wang et al., 2018b; Geng et al., 2023)	36.64	31.38
AwA2	DeViSE (Frome et al., 2013)	46.12	15.88
	OWL-based (Chen et al., 2020a)	58.90	–
	DUET (Chen et al., 2023h)	–	58.00
	OntoZSL (Geng et al., 2021a)	63.31	56.06
	DOZSL (GAN) (Geng et al., 2022)	66.36	57.62
	DOZSL (GCN) (Geng et al., 2022)	63.88	52.74
	GCNZ† (Wang et al., 2018b; Geng et al., 2023)	37.44	14.34
	GCNZ‡ (Wang et al., 2018b; Geng et al., 2023)	62.98	31.98
	FGP (Wu et al., 2023a)	79.10	43.30

RESOURCES: Several open datasets and KG resources for KG-aware ZS-IMGC have been proposed:

(i) **ImageNet** (Deng et al., 2009): A large-scale database with 14M images across 21K classes each aligned with a WordNet (Miller, 1995) entity. It leverages class hierarchies as KG-based knowledge, where the graph only contains one type of relation, i.e., *subClassOf*. From (Xian et al., 2019), a subset of 1K classes serves as seen classes, with unseen classes determined by their distance in the WordNet graph. ImageNet is widely used for ZS-IMGC benchmarking, albeit with a single-relational KG limitation.

(ii) **ImNet-A** and **ImNet-O**: Subsets of ImageNet by Geng et al. (2021a, 2023). ImNet-A contains 80 animal classes, and ImNet-O has 35 general object classes. Each comes with a KG combining multiple knowledge types, including

class attribute, class name, commonsense knowledge from ConceptNet, class hierarchy (taxonomy) from WordNet, and logical relationships such as *disjointness*.

(iii) **AwA2** (Xian et al., 2019): A coarse-grained animal classification dataset with 50 animal classes and 37,322 images, plus 85 expert-annotated attributes. Its classes align with WordNet entities for taxonomy-based KGs. Geng et al. (2023) equip AwA2 with a KG similar to ImNet-A and ImNet-O. Chen et al. (2020a) utilize OWL2 to model complex class semantics.

(iv) **NUS-WIDE** (Chua et al., 2009): A mutli-label image classfication dataset, where each image contains multiple objects. In works like (Lee et al., 2018), NUS-WIDE is accompanied by a KG with three types of label relations, including a super-subordinate correlation from WordNet, positive and negative correlations computed by label similarities such as WUP similarity.

BENCHMARKS: Single-label image classification in ZS-IMGC includes two evaluation settings: ***(i)*** **Standard ZSL:** Focuses solely on unseen class samples, using *Macro Accuracy*, which is calculated as the average of individual class accuracies (correct predictions to total samples ratio), as the metric. ***(ii)*** **Generalized ZSL (GZSL):** Evaluates both seen and unseen class samples, hence more challenging. Here, two *Macro Accuracies*, Acc_s for seen classes and Acc_u for unseen classes, are measured. The key performance indicator is the harmonic mean $H = (2 \times Acc_s \times Acc_u)/(Acc_s + Acc_u)$, ensuring a balance between the two.

In Table 2, we compile results from key methods across the three groups for benchmarks ZS-IMGC task. Incorporating diverse class semantics generally yield better ZS-IMGC results, even with basic methods. For instance, Table 2 illustrates that some mapping-based methods (Frome et al., 2013; Chen et al., 2020a), employing straightforward linear mapping functions, achieve better performance by integrating various class semantics such as attributes, hierarchy, and names, compared to GCNZ†, which relies solely on class hierarchy. Significantly, enriching GCNZ† with more class semantics, markedly enhances its effectiveness (GCNZ‡).

A.4.2 Fake News Detection

FND, also termed Rumor Detection, addresses the proliferation of misleading multimedia content on social media to ensure the dissemination of trustworthy information. Unlike standard text classification, FND challenges involve discerning falsehoods across diverse subjects. While traditional deep learning approaches in FND emphasize text, they frequently neglect the significance of visual content and background knowledge. Thus, a comprehensive integration of text, visuals, and knowledge is essential for precise FND.

KMGCN (Wang et al., 2020d) employs Entity Linking to map entities from social media posts to concepts from Probase (Wu et al., 2012) and YAGO KGs. It constructs a graph with post words as nodes and incorporates visual words from images (detected by a pre-trained object detector (Redmon and Farhadi, 2018)), weighting edges with Point-wise Mutual Information to emphasize word correlations. A GCN is then utilized to model semantic interactions, employing global mean pooling for the final binary classification of multimedia posts. Extending these insights, KMAGCN (Qian et al., 2021) integrates the visual modality through a late fusion paradigm, employing feature-level attention to more accurately delineate the interplay between visual and textual content. As a dual-consistency network, KDCN (Sun et al., 2023c) identifies inconsistencies in both cross-modal and content-knowledge aspects with Freebase as the reference KG. It finds that entities in rumor posts are more distantly connected within KGs compared to non-rumors, providing a clear marker for distinction. EmoKnow (Zhang et al., 2023d) advances COVID-19 FND by incorporating WiKiData5M (Wang et al., 2021) as an external knowledge source. It uses PLMs for text analysis, extracts emotion features, and identifies relevant linked entities, utilizing TransE (Bordes et al., 2013) for entity representation, with an MLP-based classifier to combine these multi-modal inputs.

Discussion 5 *The progression of LLMs is transforming many classification tasks into ones focused on inference and directive question-answering, emphasizing the crucial role of selecting knowledge sources. Moreover, given the frequent association of fake news with political content, their urgency of timely news highlights the need for conducting research on knowledge updating and lifelong learning in FND.*

A.4.3 Movie Genre Classification

MMGC models integrate visual, textual, and metadata information to predict movie genres, representing each genre as an element in a binary vector for

multi-genre classification of movies.

Traditional methods (Bain et al., 2020; Huang et al., 2020b) primarily rely on features extracted from images and texts alone. A recent work, IDKG (Li et al., 2023a), integrates a domain-specific MMKG created from metadata fields such as titles, genres, casts, and directors. Its motivation stems from recognizing the relational patterns in metadata, like the tendency for "*Nolan to direct science fiction movies*" or "*Emma to not often feature in comedies*". It use a translation model to merge KG embeddings with other modality features, guided by attention mechanism.

Discussion 6 *The success of a domain-specific KG largely depends on the metadata's quality and completeness, where addressing scalability is vital especially for large movie datasets. Additionally, there is scope for creating interactive and personalized Movie Genre Classification systems. By integrating user feedback and preferences, the system can be tailored to individual tastes, offering personalized genre suggestions. Techniques such as reinforcement learning and user modeling could be utilized to customize the genre classification process, thus further enhancing user experience and satisfaction.*

A.5 Supplement for KG-driven Multi-modal Generation Tasks

Some of the reasoning methods previously discussed, including those used in VQA tasks, are based on generative approaches. This section focuses on tasks where content generation is **strictly necessary** for task completion.

A.5.1 Scene Graph Generation

Introduced by Johnson et al. (2015), Scene Graphs (SGs) form a crucial data structure for scene understanding, cataloging object instances within a scene and delineating their interrelationships. These instances, ranging from people to places and objects, are described through attributes like shape, color, and pose (Chang et al., 2023). The relationships between these instances, often action-based or spatial, are expressed as *(subject, predicate, object)* triplets, paralleling the (h, r, t) and (e, a, v) triplets in KGs. Scene Graph Generation (SGG) serves as an intermediary task, unlike other multi-modal tasks with specific end goals, providing enhanced understanding and reasoning to support downstream tasks (Huang et al., 2023b; Koner et al., 2021; Yu et al., 2021).

Grasping all relationships in SGG training data is challenging, yet crucial, and leveraging prior knowledge significantly aids in effectively learning relationship representations from limited data, thereby enhancing detection, recognition, and overall accuracy of SGG. One effective approach is using **language priors**. By leveraging semantic word embeddings, these priors adjust relationship prediction probabilities, thus augmenting visual relationship identification. For example, even with infrequent occurrences in training data, like interactions between *people* and *elephants*, language priors can assist the inference of similar relationships, such as "*a person riding an elephant*", by studying more common examples like "*a person riding a horse*" (Chang et al., 2023). This also helps mitigate the long tail effect in visual relationships (He et al., 2020). Another approach involves **statistical priors**, leveraging the structural regularity inherent in visual scenes, as highlighted in (Zellers et al., 2018). These priors capitalize on typical object-relation statistical correlations, such as "*people wearing shoes*" or "*mountains being near water*".

Several works adopt the KG representation learning techniques into SGG scenario. For example, RLSV (Wan et al., 2018) uses existing SGs and images to predict new relationships between entities, targeting SG completion and blending KG embedding methods with SG characteristics in a structural-visual embedding model. Yu et al. (2022) improve zero-shot performance in SGG by constructing a KG from training set SG triples, distinguishing existing (non-zero-shot) and missing (zero-shot) edges. They train a KG Embedding model to complete the graph and fills these missing edges, thereby integrating zero-shot triples similarly to their non-zero-shot counterparts. GLAT (Zareian et al., 2020b) separates perception and commonsense into two models, training on annotated SGs with a BERT-like masking approach (akin to KG pre-training (Yao et al., 2019)) for element prediction. This method, when added to any SGG model, can rectify errors in SGs by harnessing the synergy of perception and commonsense.

Some SGG studies (Chen et al., 2019; Gu et al., 2019; Zareian et al., 2020a; Khan et al., 2022b; Chen et al., 2023f; Lu et al., 2023) also employ KGs for **triple prediction**, utilizing them to generate rich and expressive SGs. Specifically, KERN (Chen et al., 2019) leverages structured KGs to capture statistical correlations between object

pairs and relationships, boosting SGG by contextualizing and stabilizing relationship predictions to address distribution imbalances. Gu et al. (2019) utilize a knowledge-based module to identify relevant ConceptNet entities and retrieve commonsense facts, each assigned a weight indicating its real-world prevalence to filter candidate triples. Then a Dynamic Memory Network (Kumar et al., 2016) is applied for multi-hop reasoning on these facts, enabling inference of the most probable SG triples. GB-Net (Zareian et al., 2020a) views SGs as image-conditioned versions of commonsense KGs, shifting the focus from traditional entity and predicate classification to linking these two graph types. Utilizing a graph-based neural network, GB-Net iteratively propagates and refines information between and within both graphs, effectively bridging scene and commonsense knowledge. Khan et al. (2022b) enrich SGs using CSKG (Ilievski et al., 2021), a substantial commonsense KG repository. By employing graph embeddings to assess the similarity of object nodes, their approach enables graph refinement and enrichment as shown in Fig. 6. This upgrades SGG with additional information on objects' spatial proximity and potential interactions derived from external knowledge, improving higher-level reasoning and mitigating some missed or incorrect predictions made during SGG. Explicit Ontological Adjustment framework (Chen et al., 2023f) mitigates predicate biases using knowledge priors from ConceptNet and Wikidata, refining relationship detection by integrating an edge matrix from the KG into a GNN model. Tian et al. (2023) add a branch for independent label confidence estimation in SGG network, which assesses the difficulty of visual recognition. This branch balances the need for commonsense knowledge in diverse scenes, especially for relations like "*throwing*" that require supplementary knowledge compared to more straightforward spatial relations like "*sitting on*".

Discussion 7 *In the realm of SGG, KGs are instrumental in mitigating relationship bias and the long-tail phenomenon within training sets, serving as a form of refinement. However, existing SGG methods still face challenges in complex scenarios where the spatial distance between objects may be significant enough to disregard potential interactions. Enhancing scene graph integrity could be achieved by incorporating larger-scale images to recognize relationships between distantly located objects (Hossain et al., 2019). Moreover, extending SGG to identify human interactions, both in terms of object relations and social dynamics, would enrich scene comprehension and widen its practical uses, thereby aiding in the development of MMKG. Additionally, leveraging structured features from SGs in training LLMs represents a promising strategy to boost multi-modal learning, capitalizing on the combined strengths of SGs, KGs, and language priors.*

A.5.2 Image Captioning

Image Captioning (IC) (Hossain et al., 2019) is a pivotal multi-modal learning task, aiming to describe images in natural language. In IC, KGs can provide essential prior knowledge, including commonsense semantic correlations and constraints among objects, guiding the construction of semantic graphs for meaningful caption generation, even when certain elements are not visually present (Fig. 6). Furthermore, since each image in the training data typically comes with only a few ground truth captions, models often lack the cues necessary to uncover implicit intentions. KGs can significantly bridge this gap by offering essential fact-checking support.

Rule-based methods (Aditya et al., 2015; Lu et al., 2018a; Huang et al., 2020a) primarily incorporate KG knowledge into caption models through Entity Linking and symbolic rules, often supplemented by inter-concept co-occurrence scores. Aditya et al. (2015) pioneer the application of KG into IC, identifying relevant events from a KG based on detected visual concepts, then constructing a Scene Description Graph (SDG) with predefined rules, from which captions are generated using NLG tools. Lu et al. (2018a) use a CNN-LSTM model to create a template caption from the input image, followed by employing a KG-based collective inference algorithm to populate the template with specific named entities, sourced from hashtags. Instead of directly integrating semantic knowledge into the neural network layers, Huang et al. (2020a) input retrieved triples from ConceptNet during the word generation stage for next-word prediction, augmenting the probabilities of potential words identified within the semantic knowledge corpus.

Embedding-based methods (Hou et al., 2019, 2020; Li and Jiang, 2019; Mogadala et al., 2020; Zhang et al., 2021c; Zhong and Wang, 2023) typ-

ically employ networks such as GNNs or RNNs to efficiently encode retrieved knowledge as vectors, subsequently incorporating these vectors into the caption generation process. Hou et al. (2019, 2020) utilize human commonsense knowledge to support object relationship reasoning in IC, avoiding the need for pre-trained detectors. Using Visual Genome as an external KG, they map densely sampled regions from images into low-dimensional vectors, and then, guided by the KG, form a temporary semantic graph. This graph enhances GNN-based relational reasoning for captioning and iteratively refines the KG itself. CNet-NIC (Zhou et al., 2019) connects ConceptNet entries with identified image objects to enrich descriptions and infer non-explicit visual information. This method enhances the semantic depth of object recognition module outputs, integrating the embeddings of knowledge terms and image features to initialize a RNN for IC generation. Interpret-IC (Mogadala et al., 2020) selects local objects in an image based on human-interpretable rules, ensuring captions reflect only those objects of human interest. During training, entities not present in standard captions are masked to align the model with human preferences. Zhang et al. (2020) employ a chest abnormality KG with prior **chest X-ray knowledge** to support radiology report generation. In this KG, entity features are initialized with CNN-extracted features of frontal and lateral chest X-ray images, where the application of GCN mean pooling yields graph-level features that contributes to generating radiology reports. Zhao et al. (2021b) utilizes an MMKG that associates visual objects with named entities for IC, incorporating external **multi-modal knowledge** sourced from Wikipedia and Google Images. This MMKG, once processed through a GAT (Velickovic et al., 2018), feeds its final layer's output into a Transformer decoder which enables entity-aware caption generation. Nikiforova et al. (2022) propose a dataset from the Geograph project[12], including geographic coordinates of photo locations. Concentrating on **encyclopedic knowledge**, they extract facts from DBpedia and use a retriever to prioritize facts for possible caption inclusion. These knowledge triples, combined with the image and geographic context, are then utilized in an encoder-decoder IC pipeline.

[12] http://www.geograph.org.uk/

A.5.3 Visual Storytelling

Visual Storytelling (VST) transcends traditional Image Captioning by transforming a series of pictures into a cohesive narrative, demanding both the recognition of contexts within and across images and overcoming narrative monotony. KGs are crucial here, enhancing story diversity, rationality, and coherence.

KG-Story (Hsu et al., 2020) links concept terms from images across scenes using background KGs like FrameNet (Baker et al., 1998) and Visual Genome, refined by a PLM for sequential image storytelling. Yang et al. (2019a) develop a vision-aware directional encoding schema, integrating essential commonsense knowledge from ConceptNet for concept in each image. The enhanced snapshot representations, augmented with attentive knowledge, processed in a GRU-based framework for final VST. Building upon this, MCSM (Chen et al., 2021a) applies pruning rules and two concept selection modules to refine commonsense knowledge facts and facilitate sentence generation for each image using a visual-adapter-equipped BART (Lewis et al., 2020). Further, PR-VIST (Hsu et al., 2021) represents image sequences as story graphs to identify the best storyline path and develop a discriminator model for outputting story quality scores, aligning the narratives with human preferences. IRW (Xu et al., 2021) utilizes imaginary key concepts derived from each image for entity mention detection, retrieving candidate fact triples from ConceptNet to form a sub-KG. This sub-KG, along with the constructed scene and event graph for each image, is integrated using separate GCNs, adaptively contributing to the VST process. KAGS (Li et al., 2023c) involves a knowledge-enriched attention network with a group-wise semantic model for globally consistent VST guidance.

Discussion 8 *The advent of Multi-modal LLMs (MLLMs) has enriched the knowledge embedded in pre-trained models, often diminishing the need for KGs to supply coarse-grained commonsense knowledge for those IC and VST tasks. This development highlights the need for KGs offering finer-grained or specific commonsense knowledge to address model hallucination issues. Moreover, for VST task, maintaining coherence between pictures and scenes is essential, where KGs are vital for linking disparate scenes and enriching scene transitions with background knowledge. Several methods have innovated with data-centric KG enhance-*

ments, such as deriving background KGs from story collections in training corpora (Hsu et al., 2021), or creating event graphs through image selection from the training set that resemble the query image, subsequently using Information Extraction tools to construct events for each sentence associated with an image (Xu et al., 2021). While these strategies are pioneering, they introduce challenges in ensuring equitable model comparisons due to varied dependencies on external knowledge sources, suggesting the need for separate evaluation of such data-centric methods.

A.5.4 Conditional Text-to-Image Generation

Conditional Text-to-Image Generation (cIG) aims to transform textual descriptions into visually realistic images, where KGs could supply detailed prior knowledge and commonsense elements not originally present in the datasets. LeicaGAN (Qiao et al., 2019) establishes a shared semantic space that enables text embeddings to convey visual information, by integrating a text-image encoder for semantic, texture, and color understanding, alongside a text-mask encoder for shaping layout through segmentation masks. During the image imagination phase, it merges the outputs of these encoders with added Gaussian noise to enhance diversity. Here, a cascaded attentive generator produces detailed and realistic images, ensuring semantic and visual coherence through adversarial learning. Many following works (Cheng et al., 2020; Jun Cheng and Fuxiang Wu and Yanling Tian and Lei Wang and Dapeng Tao, 2022; Liu et al., 2023b) treat image-caption pairs in training sets as KB entries, enriching captions by selecting and refining relevant items from this KB, thereby aiding in feature extraction and enabling more accurate cIG. Concretely, KnHi-GAN (Ge et al., 2021) and AttRiGAN (Zhu et al., 2021) present a Knowledge-enhanced Hierarchical GAN, employing a KG to enrich text descriptions for detailed generative input. This task-specific KG is constructed from training sample attributes, formatted in RDF triples (Geng et al., 2021a, 2023). For 3D cIG, T2TD (Nie et al., 2023) involves a text-3D KG that correlates text with 3D shapes and textual attributes, utilizing these elements as prior knowledge. During 3D generation, it retrieves the knowledge based on text descriptions and employs a causal module to select shape information relevant to the text.

Discussion 9 *While metrics like the Inception score (Salimans et al., 2016) and R-precision (Xu et al., 2018) are commonly used for evaluating the diversity of generated images and the semantic consistency between input text and generated images, current evaluation methods for generated images still lack critical assessment at the knowledge and commonsense level (Huang et al., 2023b). Bridging this gap presents a critical direction for future research.*

A.6 KG-driven Multi-modal Retrieval Tasks

Definition 2 ***KG-aware Retrieval*** *aims to utilize textual descriptions ($x^{\mathbb{l}}$) for ranking similar visual images ($x^{\mathbb{v}}$), or vice versa, including the sorting and retrieval of all relevant images or region proposals within an image. Utilizing a background KG $\mathcal{G}$, this approach transcends mere appearance-based retrieval by incorporating non-visual attributes, striving for a human-level semantic understanding, especially in scenarios lacking precise targets.*

A.6.1 Cross-Modal Retrieval

Cross-Modal Retrieval (CMR) focuses on fetching data across different modalities, such as images, text, audio, or video, in response to a query from another modality. Specifically, this section explores Image-Text Retrieval, aiming to identify semantically similar instances across visual and textual modalities.

Image-Text Matching (ITM) vs. Image-Text Retrieval (ITR): ITM and ITR are closely related yet differ mainly in their application: ITM evaluates relevance between an image and text, often used in image-caption correspondence (Huang et al., 2023b; Gómez-Pérez and Ortega, 2019), while ITR focuses on finding relevant matches in larger datasets based on textual or visual queries, crucial for visual search engines, digital asset management, and automated content generation (Cao et al., 2022a). Both ITM and ITR leverage similar underlying technologies, metrics, and datasets such as Flickr30k (Young et al., 2014) and MSCOCO (Lin et al., 2014), which feature extensive labeled images with captions. In cross-modal pre-training, ITM serves as a foundational task, honing the model's ability to semantically correlate images and text, thereby improving its effectiveness in ITR (Zhang et al., 2021b; Li et al., 2022a, 2023b). This pre-training ranges from coarse-grained matching (assessing general semantic relatedness) to fine-grained matching (aligning specific image regions with text). Such granularity

enhances the nuanced understanding and retrieval capabilities for pre-trained models, bridging the gap between general-purpose models and the specific demands of ITR tasks.

Early CMR research often overlook long-tail and occluded semantic concepts in images (Yang et al., 2021; Cao et al., 2022a). Recent advancements (Shi et al., 2019; Wang et al., 2020a, 2022a; Li et al., 2023d; Yang et al., 2023a) tend to fix this by leveraging knowledge from frequently co-occurring concept pairs in Visual Genome's scene graphs (Krishna et al., 2017) or image captioning corpus. They create Scene Concept Graphs (SCGs) using heuristic or rule-based tools such as language parsers (Anderson et al., 2016), aiming to capture fine-grained details. Shi et al. (2019) initially identify broad concepts, further refined into detailed ones via SCG's co-occurrence relationships, followed by a concept prediction module for accurate labeling. EKDM (Yang et al., 2023a) employs an iterative concept filtering module that progressively incorporates candidate concepts into a static global representation in a dynamic manner, which uses the significance scores of these concepts to set the fusion order, integrating higher-scored concepts first. CVSE (Wang et al., 2020a, 2022a) utilizes a GNN for semantic correlation propagation in SCG, enriching concept representations with commonsense knowledge via weighted embedding summation. A confidence scaling function is introduced to mitigate long-tail distribution challenges. CSRC (Li et al., 2023d) further employs a multi-head self-attention mechanism to selectively focus on deeper conceptual emphasis, while MACK (Huang et al., 2022) eliminates the need for paired domain data during training.

Note that the background KGs in these works are typically derived from large-scale multi-modal datasets, rather than directly utilizing public KGs. However, **a reliance on mere word co-occurrences for entity similarities can be misleading**, like wrongly linking "*man*" and "*dog*" due to frequent co-occurrences. Utilizing WordNet's noun hierarchies helps distinguish such entities. Additionally, MMKGs could address this by capturing inter-modal co-occurrence relations, like temporal, causal, and logical connections. For instance, "*washing*" with "*tap*" or "*cutting*" with "*knife*" in image-text pairs enhances semantic understanding across modalities. Building on this perspective, Fig. 11 illustrates MMKG-based approach MKVSE (Feng et al., 2023), which enhances image-text semantic connections, especially for images with indirect textual descriptions. It scores intra- and inter-modal relations in MMKGs using WordNet path similarity (calculated by NLTK (Bird et al., 2009)) and co-occurrence correlations, improving ITR through GNN-based embeddings. Moreover, Yang et al. (2023a) focus on a common limitation in visual concept modeling, where **varying spatial locations are often inaccurately linked by fixed relationships**, like "*man-on-bike*" for any proximity of "*man*" and "*bike*" in an image. They spatial information from a geometric graph (Monti et al., 2017) to discern spatial relations between image regions and employ a location CNN model to refine visual-semantic representations. EGE-CMP (Dong et al., 2023) is a entity-graph enhanced cross-modal pre-training framework that leverages entity knowledge extracted from captions instead of human labeling. It focuses on learning instance-level feature representations by infusing real semantic information into visual-text alignment, improving text-image cross-modal alignment.

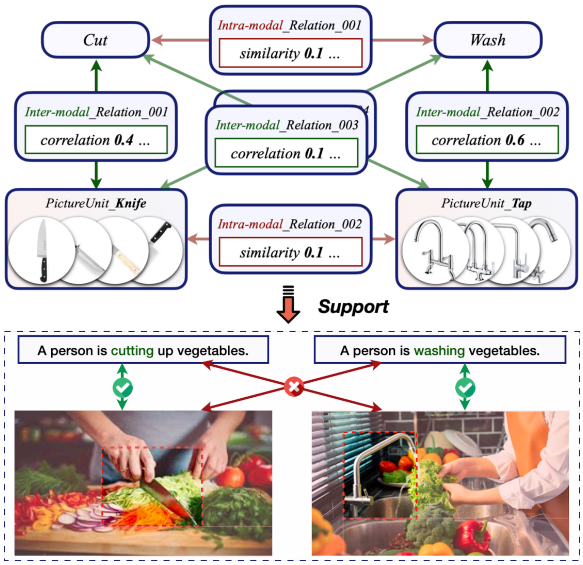


Figure 11: We illustrates the MMKG-supported Image-Text Retrieval process (Feng et al., 2023). For simplicity, all URI prefixes and certain relations (*sourceImg* and *targetImg*) from the *PictureRelation* (*Inter-modal_Relation* and *Intra-modal_Relation*) entity are omitted. This entity's values indicate 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.

Discussion 10 *Current VLMs face challenges in fine-grained cross-modal semantic matching. Wang et al. (2023b) tackle this issue by using contrastive*

learning for aligning entities from Visual Genome in ITR, enhancing cross-modal sensitivity with entity masking. We note that a shift towards knowledge-guided strategies rather than relying solely on co-occurrence in VLM training could significantly improve retrieval and matching of fine-grained, long-tail objects, potentially leading to advanced semantic grounding (Chen et al., 2023h) and wider applications. However, only limited studies (Feng et al., 2023) have considered the role of external knowledge like WordNet's semantic structures. Besides, as discussed in §A.3.1, various types of KGs, including trivia, commonsense, scientific, and situational knowledge, offer unique and complementary insights for reasoning processes. But the prevalent focus on co-occurrence information captures a fraction of commonsense knowledge. Looking ahead, exploiting long-tail knowledge from diverse large-scale KBs holds significant potential for enhancing models' generalization capabilities across various domains and real-world scenarios.

A.6.2 Visual Referring Expressions & Grounding

This section revisits KG-aware approaches in Visual Referring Expressions (also known as Phrase Grounding or Referring Expression Comprehension) and Visual Grounding. While CMR typically entails matching across diverse textual and visual contexts, VRE and VG focus on aligning fine-grained features within specific textual-visual pairs. From a certain point of view, these tasks are akin to adding an extra step of grounding answers in the conventional KG-based VQA, as illustrated in Fig. 9.

Visual Referring Expressions (VRE) vs. Visual Grounding (VG): VRE and VG (Qiao et al., 2021) integrate linguistic and visual information, differing in focus (Qiao et al., 2021): VRE identifies and localizes a specific image region that corresponds to a given textual expression, typically involving a detailed description of one object. Conversely, VG is about localizing various object regions linked to multiple noun phrases in a sentence, aiming to establish fine-grained alignment between vision and language. Despite these differences, both tasks require deep semantic language interpretation and manage ambiguities inherent in natural language and visual perception, relying on extensive annotated datasets. The line between VRE and VG often blurs in research, with some approaches (Yang et al., 2019b; Sibei Yang and Guanbin Li and Yizhou Yu, 2021; Tang et al., 2023) merging their key aspects: VRE's precise object localization and VG's broad contextual analysis.

KAC Net (Chen et al., 2018) utilizes the knowledge from pre-trained fixed category detectors, essential for selecting relevant proposals and ensuring visual consistency, to filter out unrelated proposals in VG progress. Shi et al. (2022) tackle **zero-shot VRE**, where visual examples of queried object categories in the test set are not shown in the training set (i.e., open-vocabulary scene); they achieve this by dynamically building MMKGs using commonsense knowledge from WordNet and ConceptNet, combined with situational knowledge from Visual Genome. Query-derived entities, detected objects, and predefined relationships are integrated into these MMKGs, employing GCN for node representation and defining eight spatial relations to assist localization of noun phrases.

The **KB-Ref** dataset (Wang et al., 2020b) emphasizes commonsense knowledge, with its construction process inspired by the F-VQA dataset (Wang et al., 2018a), which involves creating a commonsense KG. Concretely, volunteers craft referring expressions for queried objects based on facts from this KG, deliberately avoiding the use of specific object names. Building upon the KB-Ref dataset, ECIFA (Wang et al., 2020b) introduces a multi-hop facts attention module from the KG and a matching module that utilizes expression-object scores for accurate grounding; CK-Transformer (Zhang et al., 2023f), leveraging the UNITER (Chen et al., 2020c) as its backbone, selects top-K retrieved facts from the KG for a given expression and visual region candidates, encoding these into multi-modal features to compute matching scores for each candidate. Bu et al. (2023) observe that knowledge-based Referring Expressions often consist of two segments: visual segments (e.g., "*on the sofa*" in Fig. 9), interpretable directly from visual content like color and shape, and knowledge segments (e.g., "*used for sleeping*" in Fig. 9), requiring additional information beyond visuals like function and non-visual attributes. To mitigate similarity bias, they introduce the SLCO network, which uses knowledge segments for category retrieval and visual segments for object grounding.

The **SK-VG** dataset (Chen et al., 2023g) targets at scene knowledge-guided VG, using movie scene images from the VCR dataset (Zellers et al., 2019). Designed to promote reasoning beyond mere image

content, SK-VG employs a detailed two-stage annotation process: firstly, generating story descriptions for each image, and secondly, crafting referring expressions tied to these stories and images, accompanied by object bounding box annotations. These annotations are crafted to ensure knowledge relevance to the scene context, uniqueness for accurate object identification, and diversity in both lexical use and objects. Chen et al. (2023g) further provide two benchmarking algorithms: a one-stage approach that embeds knowledge into image features prior to query interaction, and a two-stage method that extracts features from images and text, subsequently employing structured linguistic data for computing region-entity similarity.

Discussion 11 *A good VRE and VG system can benefit various downstream tasks such as VQA, CMR, and IMGC. Chen et al. (2023h) develop a cross-modal semantic grounding network for ZS-IMGC, aimed at disentangling semantic attributes from images via a self-supervised method. This technique bridges knowledge from PLMs to visual models without needing region-attribute supervision. By leveraging AWA2-KG (Geng et al., 2023) for fine-grained labeling, it connects species to their attributes (e.g., "zebra" to "striped") and uses KG serialization to blend structured knowledge into cross-modal grounding. The network also incorporates attribute-level contrastive learning to tackle attribute imbalance and co-occurrence, thus refining the distinction of fine-grained visual features across images from both seen and unseen classes. This highlights the value of KGs in Visual Grounding tasks, serving as a natural knowledge organizer and a conduit for transferring VG principles to related tasks without specialized annotations.*

A.7 Supplement for KG-driven Multi-modal Pre-training

In this section, our primary focus is on pre-training definitions related to Transformer-based models, aligning with the current mainstream discourse in AI community. Other paradigms, such as Poincaré embedding pre-training (Xu et al., 2020), are not covered in this discussion.

We highlight that multi-modal reasoning and generation tasks often require an extensive range of specialized knowledge, typically involving long-tail information that goes beyond everyday 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. A significant challenge arises in tasks requiring visual common sense, where models may falter due to limited cross-modal alignment capabilities, leading to multi-modal hallucinations as evidenced in Fig. 7. Recent works (Zha et al., 2023) demonstrate that MMKGs can effectively bridge this gap, enhancing the potential of multi-modal methods and offering a robust solution to multi-modal hallucinations in the era of LLMs. Specifically, Zha et al. (2023) introduce M^{2}ConceptBase, a multi-modal conceptual MMKG. They develop a pipeline using M^{2}ConceptBase to improve knowledge-based VQA performance by retrieving multi-modal concept descriptions and crafting instructions to refine answers with MLLMs.

Structure Knowledge aware Pre-training. The integration of structured knowledge into multi-modal content understanding has gradually gained momentum, drawing inspiration from advancements in the NLP field. KM-BART (Xing et al., 2021) adapts the BART (Lewis et al., 2020) model to multi-modal tasks by incorporating a pre-trained visual feature extractor. It tackles knowledge-based commonsense generation by using COMET (Bosselut et al., 2019) to augment image-caption datasets with commonsense context. The enriched datasets, combined with a next-token prediction target, empower KM-BART to deduce events and character intentions from image-text pairs. ERNIE-ViL (Yu et al., 2021) incorporates Scene Graph (SG) knowledge into a VLM, enhancing visual scene comprehension by adding SG completion and prediction tasks (covering objects, attributes, and relationships) during its multi-modal pre-training stage. ROSITA (Cui et al., 2021) strengthens semantic alignments across visual and language modalities by employing a unified SG shared between the input image and text. Existing VLMs often struggle with Image-Text Matching tasks that demand an understanding of reversed roles or actions, evident in scenarios like "*An astronaut rides a horse*" versus "*A horse rides an astronaut*" (refer to § A.6.1). To tackle this, Structure-CLIP (Huang et al., 2023b) improves structured multi-modal representation learning by leveraging SGs to generate semantic

negative examples.

Knowledge Graph aware Pre-training. KG-Transformer (Zhang et al., 2023b) is pre-trained on KGs including WN18RR (Sun et al., 2019b), FB15k-237 (Toutanova and Chen, 2015), and CoDEx (Safavi and Koutra, 2020), with pre-training objectives like masked relation/entity prediction and entity pair prediction. This model can be applied to ZS-IMGC, framing the task as determining the match score between input images and target classes. It undergoes fine-tuning using AwA-KG (Geng et al., 2023), with a pre-trained ResNet serving as the vision encoder and image representations further transformed through a trainable matrix.

Discussion 12 *Current KG-equipped VLMs primarily use triple contexts to enhance multi-modal data, with a few examples, like KGTransformer (Zhang et al., 2023b), incorporating KG's structural information into pre-training. However, its application is limited to Zero-shot Image Classification, using a uni-modal approach during pre-training. Future research in this domain can focus on four key areas: Firstly, scaling up KG to exploit its rich knowledge and structural traits, rethinking the long-tail phenomenon in multi-modal pre-training data and expanding the knowledge scope to involve world knowledge. Secondly, the integration of MMKGs, which are further discussed in § 4. Third, exploring unique pre-training paradigms suited for (MM)KGs to fully harness the value of structured knowledge in multi-modal pre-training. Fourth, extending to more downstream tasks to align with the latest advancements in AGI, utilizing MLLMs (Liu et al., 2023c).*

A.8 Supplement for Future Directions

Focusing solely on the benefits that traditional KGs bring to multi-modal tasks can be inherently limiting due to the restricted scope of knowledge captured in single-modality KGs. In evaluating KG-aware multi-modal tasks, it's crucial to discern the unique advantages of multi-modal knowledge, especially compared to large-scale textual or multi-modal corpora. Specifically for image modalities, MMKGs can be categorized into two types: A-MMKGs where images serve as entity attributes, and N-MMKGs where images are independent entities with their own relationships (Zhu et al., 2022). A pivotal question is whether structured (MM)KGs offer irreplaceable benefits that maximize their potential. Additionally, we should consider whether non-LLM models augmented by (MM)KG can rival or outperform MLLMs in specific tasks, providing compelling reasons to support the future development.

A.8.1 KG4MML Tasks.

Multi-modal Content Generation. Current applications of MMKGs in multi-modal content generation are quite limited. Most existing efforts only integrate KGs to supplement additional context beyond datasets or to connect different visual scenes. Future developments should aim for larger, more detailed MMKGs to employ multi-modal structural data in training, fostering more controlled and logically coherent generation and mitigating hallucinations.

Multi-modal Task Integration. Different domains currently evolve independently with limited cross-interaction. In Cross-Modal Retrieval (CMR), (MM)KGs are widely employed for information enhancement, whereas in knowledge-based VQA, the focus is mainly on dense vector retrieval and modality conversion techniques. This highlights the potential for future advancements like integrating KG-based CMR methods into KG-based VQA. In a similar vein, generation tasks can enhance retrieval, reasoning, and discrimination, with knowledge-enhanced discrimination tasks playing a key role in refining answers for other tasks. As knowledge-intensive multi-modal tasks gain prominence, merging these distinct domains with (MM)KG at the core will becomes crucial.

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:

- **Non-Uniform Organization and Ontology of MMKGs:** Current MMKGs, lacking a stan-

dardized 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., 2023c; Zha et al., 2023), with commonsense and scientific related MMKGs (Wang et al., 2023c; 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., 2023c).

- **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.
- **Data Timeliness and Completeness Issues in MMKGs** heightens the risk of multi-modal hallucinations.
- **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.

- **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., 2023e; 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., 2023h).
- **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.
- **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 Multi-modal Knowledge Graph Construction, Multi-modal Entity Alignment, can enhance domain integration. There are already some attempts in this direction (Pahuja et al., 2024).

A.8.2 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.

1. **Fine-Tuning**:
   - MMKGs provide a rich source of structured multi-modal data for Supervised Fine-Tuning (SFT) of Multi-modal LLMs (MLLMs), Training techniques effective for MMKGs in VLMs can also be applied to MLLMs, especially in domain-specific applications (Zheng et al., 2024; Zhang et al., 2023c).
   - Leverageing self-instructing techniques to autonomously evolve and generate multi-grained, multi-modal instructional data (Wang et al., 2023d; Xu et al., 2023a; Du et al., 2023; Yona et al., 2024)
   - 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.
2. **Hallucination**:
   - As LLMs rapidly advance, the risk of generating seemingly authentic but factually inaccurate web content is increasing. This phenomenon, known as *hallucination* (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 (Liu et al., 2024).

- Hallucination in MLLMs: (Huang et al., 2023a; Tong et al., 2024; Liu et al., 2024)
- To combat this, LAMM (Yin et al., 2023b) 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., 2023c). Developing robust hallucination detectors (Chen et al., 2023e; 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., 2023a).

3. **Agent**:
   - Multi-agent Collaboration (Xu et al., 2023b; 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., 2023e; 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., 2023a). For example, Sun et al. (2023a) 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.
   -

4. **Retrieval Augmented Generation**:
   - 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., 2023f). 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).