Natural Language Understanding and Inference with MLLM in Visual Question Answering: A Survey

JIAYI KUANG, Sun Yat-Sen University, Shenzhen, China
YING SHEN, Sun Yat-Sen University, Shenzhen, China
JINGYOU XIE, Sun Yat-Sen University, Shenzhen, China
HAOHAO LUO, Sun Yat-Sen University, Shenzhen, China
ZHE XU, Sun Yat-Sen University, Shenzhen, China
RONGHAO LI, Sun Yat-Sen University, Shenzhen, China
YINGHUI LI, Tsinghua University, Shenzhen, China
XIANFENG CHENG, Sun Yat-Sen University, Shenzhen, China
XIKA LIN, Department of Computer Science, Worcester Polytechnic Institute, Worcester, United States
YU HAN, Sun Yat-Sen University, Shenzhen, China

Visual Question Answering (VQA) is a challenge task that combines natural language processing and computer vision techniques and gradually becomes a benchmark test task in multimodal large language models (MLLMs). The goal of our survey is to provide an overview of the development of VQA and a detailed description of the latest models with high timeliness. This survey gives an up-to-date synthesis of natural language understanding of images and text, as well as the knowledge reasoning module based on image-question information on the core VQA tasks. In addition, we elaborate on recent advances in extracting and fusing modal information with vision-language pretraining models and multimodal large language models in VQA. We also exhaustively review the progress of knowledge reasoning in VQA by detailing the extraction of internal knowledge and the introduction of external knowledge. Finally, we present the datasets of VQA and different evaluation metrics and discuss possible directions for future work.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Computer vision;*

Additional Key Words and Phrases: Visual question answering,multimodal representation and reasoning,multimodal large language models

Authors' Contact Information: Jiayi Kuang, Sun Yat-Sen University, Shenzhen, China; e-mail: kuangjy6@mail2.sysu.edu.cn; Ying Shen (Corresponding author), Sun Yat-Sen University, Shenzhen, China; e-mail: sheny76@mail.sysu.edu.cn; Jingyou Xie, Sun Yat-Sen University, Shenzhen, China; email: xiejy73@mail2.sysu.edu.cn; Haohao Luo, Sun Yat-Sen University, Shenzhen, China; e-mail: luohh5@mail2.sysu.edu.cn; Zhe Xu, Sun Yat-Sen University, Shenzhen, China; e-mail: xuzh226@mail2.sysu.edu.cn; Ronghao Li, Sun Yat-Sen University, Shenzhen, China; e-mail: lirh56@mail2.sysu.edu.cn; Yinghui Li, Tsinghua University, Shenzhen, China; e-mail: liyinghu20@mails.tsinghua.edu.cn; Xianfeng Cheng, Sun Yat-Sen University, Shenzhen, China; e-mail: chengxf6@mail2.sysu.edu.cn; Xika Lin, Department of Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts, United States; e-mail: xikalin@gmail.com; Yu Han, Sun Yat-Sen University, Shenzhen, China; e-mail: hanyu25@mail.sysu.edu.cn.

ACM 0360-0300/2025/03-ART190
https://doi.org/10.1145/3711680

ACM Reference Format:
Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural Language Understanding and Inference with MLLM in Visual Question Answering: A Survey. *ACM Comput. Surv.* 57, 8, Article 190 (March 2025), 36 pages. https://doi.org/10.1145/3711680

1 Introduction

1.1 What is Visual Question Answering?

Visual Question Answering (**VQA**) has emerged as a significant task in the intersection of computer vision and **natural language processing** (**NLP**). The goal of VQA is to predict an answer A to a question Q based on visual information V, formalized as $A = f(Q, V)$, where f represents the model function [5]. Visual inputs may include images of landscapes, people, or videos, and questions can range from multiple-choice to open-ended formats.

A VQA model typically consists of the following steps. First, features are extracted from visual and textual information respectively. Then, intra-modal and inter-modal representations are learned by aligning and fusing the features. Finally, the obtained image-question knowledge representation is predicted to complete the question answering task. We summarize these natural language and image understanding and inference in Figure 1.

The VQA task, introduced by Agrawal et al. [5] with the VQA v1 dataset comprising 614,163 curated question-answer pairs based on MS COCO images, marked the start of a crucial research direction. Early efforts primarily focused on direct multimodal understanding, leveraging deep learning methods. Visual features were extracted using models like VGG-Net [43] and Faster-RCNN [130], while textual features were processed through LSTM [5] and GRU networks. The advent of attention mechanisms significantly advanced the field. Frameworks such as stacked attention [174] and co-attention [99] improved visual-textual feature integration by learning inter-modal dependencies. Additionally, **Graph Neural Networks** (**GNNs**) have been employed to model complex relationships via multimodal heterogeneous graphs [26].

More recently, the Transformer architecture [157] has driven substantial progress in VQA. Visual-language pre-trained models like BLIP [78] utilize self-attention and cross-attention mechanisms for enhanced multimodal fusion, enabling breakthroughs in zero-shot VQA. Emerging **multimodal large language models** (**MLLMs**), including Flamingo [3], BLIP-2 [77], and InternVL [19], excel in open-ended and zero-shot question answering scenarios.

While early VQA models focused on direct feature extraction and alignment, recent advances have shifted toward knowledge-based reasoning, which goes beyond perception to deeper cognitive understanding. Some approaches explore how to transform fragmented internal knowledge into structured knowledge. Knowledge can be represented as the form of knowledge triplets and applied to one-hop or multi-hop reasoning [26]. Other approaches form a joint image-question-knowledge representation by introducing an external knowledge base [48] or performing passage retrieval [37]. Then, this joint representation is utilized for reasoning. Compared with the previous work, the understanding of the knowledge representation coming from images and texts at this stage is more in-depth. In addition, Multimodal large language models now integrate sophisticated reasoning methods such as instruction tuning and Chain-of-Thought prompting to improve answer accuracy. In this respect, VQA is a comprehensive task that bridges computer vision and NLP. On the one hand, computer vision aims to teach machines how to see, working on ways to acquire, process, and understand images. NLP, on the other hand, is a field concerned with enabling interactions between computers and humans in natural language, which not only aims to teach machines how to read but also pays attention to the thought process of question answering. It is worth noting that natural language generation methods play an important role in VQA, since it

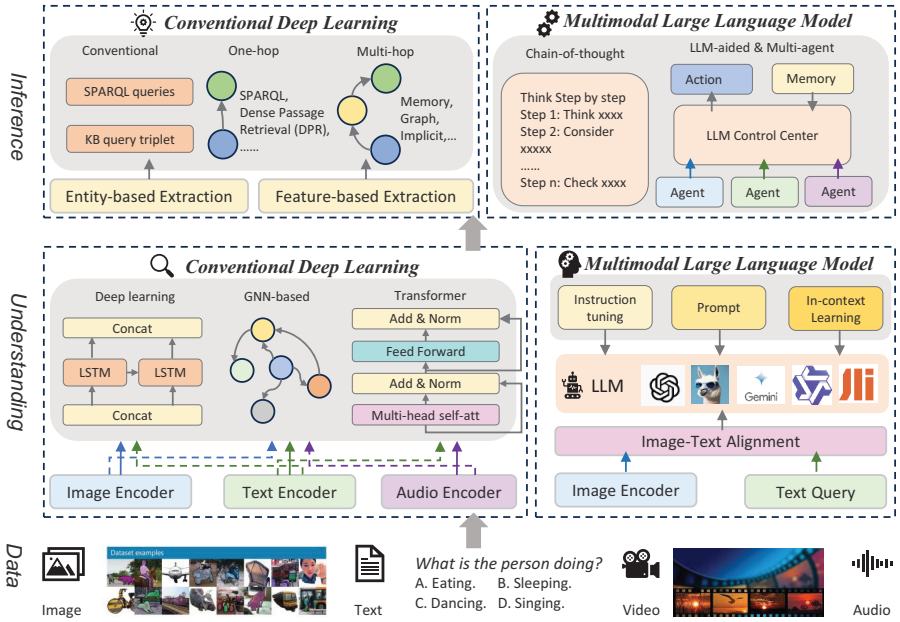


Fig. 1. The data used in VQA tasks, with the conclusion of the understanding and inference methods from conventional models to multimodal large language models.

has a non-negligible role in the answer generation process, especially contributing to assisting the model to achieve better results in open-ended question answering.

VQA bridges computer vision and NLP, integrating visual understanding–teaching machines to process and interpret images–with natural language generation, which supports nuanced question-answering processes. Notably, these developments have propelled VQA's utility across diverse applications, including aiding visually impaired individuals, improving image retrieval, autonomous driving, medical diagnostics, and conversational AI [10, 87, 145]. VQA also extends into visual dialogue [105], where systems engage with conversational context, further demonstrating its potential for real-world problem-solving.

1.2 Comparison with Other Survey Works

Over the years, VQA has garnered significant attention in the research community, leading to the publication of numerous high-quality surveys. These surveys provide valuable insights for both beginners and experienced researchers by outlining challenges, recent trends, and open problems. Table 1 highlights some of the prominent generalized surveys in the field.

In 2017, Wu et al. [164] provide a foundational review, offering an overall definition of VQA task types and a comprehensive overview of existing models. However, their survey did not cover more advanced VQA approaches. That same year, Kafle et al. [62] review VQA tasks with a focus on datasets and evaluation metrics, discussing the challenges and limitations of existing methods. Zhang et al. [185] later review information fusion methods in VQA, categorizing them into two-channel and multi-channel fusion strategies. While this work advanced the understanding of multimodal fusion, it provides limited insights into other critical components of the task. By 2020, Srivastava et al. [146] highlight the latest applications of deep learning methods in VQA, offering a detailed analysis of model performance. Patel et al. [123] further contribute by focusing on video-based question answering, emphasizing temporal information processing. More recently,

Table 1. Some of the Prominent Generalized Long Surveys along with the Published Year, Topics, Challenges, and Key Contributions

Surveys	Year	Topic	Challenges	Contributions
Wu et al. [164]	2017	Visual Question Answering, Knowledge Base	Question form is unknown, Visual and textual comprehension, Lack of knowledge	First comprehensive overview of the field, Definition and classification of the task, In-depth analysis of the question/answer pairs
Kafle et al. [62]	2017	Image Understanding, VQA datasets	Solving a wide range of CV tasks, Dataset bias, VQA algorithm	Compare VQA with other computer vision tasks, Exploring whether current VQA benchmarks are suitable for evaluating
Zhang et al. [185]	2019	ImageQA and VideoQA, Information Fusion	Fusion of semantic information of text and vision, temporal relationship in VideoQA	Abstract fusion framework that can fit the majority of existing VQA models, Two-channel fusion and multi-channel fusion, Clear organization on the proposed fusion techniques
Srivastava et al. [146]	2021	Deep learning in VQA, Robust datasets	Real-life Datasets, Prior knowledge bias	Cover major datasets published for validating the Visual Question Answering task, Discuss state-of-the-art architectures and compare results
Patel et al. [123]	2021	Video Question Answering, Temporal reasoning	Collection of Video-based QA dataset, Video content must have varied actions	Review a number of methods and datasets for VideoQA
Faria et al. [22]	2023	Language bias for VQA, Scene TextVQA, OOD data reasoning	Generalization in VQA, Zero-shot VQA, Consolidated VQA benchmark	Discuss the steps involving VQA task, Introduce the most recent and significant works comprising strategies for VQA pipeline
Md.F. Ishmam et al. [55]	2024	vision language pre-training	encompass traditional VQA and VLP-based methods	Introduces a detailed taxonomy to categorize VQA, Highlights the recent trends, challenges, and scopes for improvement for more domain such as multimodalQA
Ma et al. [108]	2024	Robust VQA, dataset bias	Poor performance in out-of-distribution dataset of VQA	Overview of in-distribution and out-of-distribution datasets, Typology that presents existing debiasing methods

Faria et al. [22] explore the language bias problem in VQA, providing detailed analyses of scene-text datasets and strategies to address bias. Barra et al. [10] and Singh et al. [145] offered shorter surveys that summarized recent advances in neural network-based models, pre-trained language models, and their applications. Additionally, Ishmam et al. [55] presented a taxonomy of vision-language pretraining strategies and their relevance to VQA, while Ma et al. [108] addressed dataset bias and proposed debiasing methods to enhance VQA robustness.

Compared with previous work, our survey provides an up-to-date overview of VQA development, with a particular focus on the latest models and their timeliness. We address knowledge reasoning techniques applied in recent years and the multimodal large language models in few-shot VQA, which have been underexplored in prior surveys.

1.3 Contribution of this Survey

In this article, we give details of the processing models, relevant datasets, and evaluation methods for the VQA task. We define the framework paradigm and taxonomy of VQA task in Figure 2, including natural language understanding of image and text and natural language inference. The main contributions of our article are as follows:

(1) To give an up-to-date synthesis of VQA paradigm, including natural language understanding of images and text (perceptual ability), as well as the natural language inference module based on image-question information (cognitive ability) on the core VQA tasks, as shown in Figure 2;
(2) To elaborate on the latest progress of visual-language pre-training models, graph neural network models, and multimodal large language models in VQA for natural language understanding of images and text;
(3) To highlight the progress of natural language inference in VQA by detailing the knowledge reasoning and multimodal large language model reasoning methods.

The rest of this survey is organized as follows: In Section 2, we discuss natural language understanding for images and text, focusing on improved extraction, embedding, and multimodal fusion methods. Section 3 reviews large language models in zero-shot VQA, introducing three paradigms for their application. In Section 4, we explore natural language inference, including knowledge sources, acquisition, and reasoning processes, along with multimodal large language

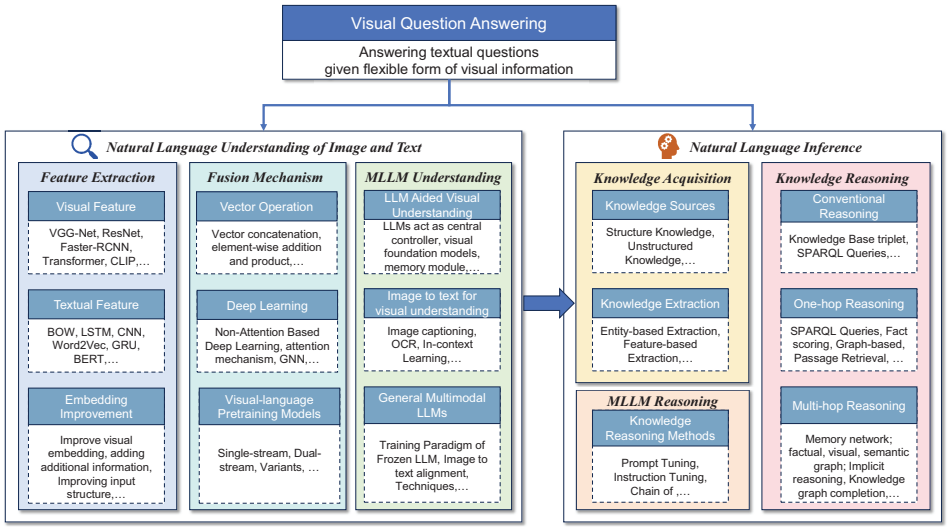


Fig. 2. Taxonomy graph of VQA task.

model reasoning methods in VQA. Finally, we discuss VQA challenges and propose future research directions in Section 6, concluding in Section 7.

2 Comprehension of Image and Text In VQA

2.1 Feature Extraction

The majority of VQA models require modal feature extraction prior to answering questions, which can be used for subsequent multimodal feature fusion to eliminate the gap between modals.

2.1.1 Visual Feature Extraction. For visual feature extraction, on the one hand, **convolutional neural networks (CNNs)** (e.g., VGG-Net [144], GoogleNet [150] and ResNet [47]) are often used to extract global image features in early VQA tasks. VGG-Net [144] increases the convolutional layers to 19 and replaces 5×5 with 3×3 convolutions, reducing parameters and enhancing non-linear mapping for improved expressive ability. ResNet [47] introduces residual blocks to mitigate gradient issues in deeper networks, weakening strong connections. On the other hand, **Visual Transformer (ViT)** [29] applies the Transformer [157] framework to extract image features and compute attention map of the image by attending each pixel to every other pixel. Based on this, CLIP [126] shows the strong performance to extract features on pixel-level, and further becomes a widely-used image encoder to assistant the vision-language pretraining models and multimodal large language models with pixel-level image understanding ability. Additionally, visual deep MLP [95] has emerged as a shift to novel visual understanding paradigm.

Using global image features to perform VQA task weakens the relationship between tasks and objects in the image, so numerous models prominent the task-relevant regions by extracting region-based image features. Actually, it is a spatial-attention mechanism to extract finer-grained features. They first select the regions of interest to the task and input them into CNNs to extract region features. Specifically, there are two ways to extract region-based features. One is based on the uniform grid [58]. By dividing image into uniformly sampled grids, the region features corresponding to each grid can be achieved after inputting them into CNNs. And, the relevance weight of each grid feature is determined by the task. Another way is based on region proposal, which applies object recognition techniques to generate bounding boxes for the image, then inputs them

Table 2. The VQA Models with Different Published Year, Architecture, and Dataset

Models	Year	Architecture			Datasets
		Visual Feature	Textual Feature	Fusion Strategy	
IBOWIMG	2015	GoogLeNet	BoW	Vector Concatenation	DAQUAR, VQA1.0, MS COCO
ABC-CNN	2015	VGG -Net	LSTM	Element-Wise Addition, CNN	DAQUAR, VQA1.0, COCO-QA
SAN	2016	VGG -Net	LSTM	Element-Wise Addition, Attention	VQA1.0, COCO-QA
Full-CNN	2016	VGG -Net	CNN	CNN	DAQUAR, COCO-QA
AMN	2020	VGG -Net	Word2Vec	Attention	MovieQA
Marioqa	2017	C3D	GRU	Attention	CLEVER, MovieQA
MLB	2016	ResNet	GRU	Bilinear Pooling Fusion	VQA1.0
MCB	2016	ResNet	LSTM	Bilinear Pooling Fusion	VQA1.0, VQA2.0, Visual7W
Pixel-BERT	2020	ResNet	BERT	Transformer	VQA1.0, VQA2.0
SOHO	2021	ResNet	BERT	Transformer	VQA1.0, VQA2.0
LXMERT	2019	Faster-RCNN	BERT	Cross-Modal Transformer	A-OKVQA, GQA, VizWiz
ViLBERT	2019	Faster-RCNN	BERT	Cross-Modal Transformer	A-OKVQA, VQA2.0
Oscar	2020	Faster-RCNN	BERT	BERT	VQA2.0
ConceptBert	2020	Faster-RCNN	BERT	Transformer	A-OKVQA, VQA1.0
MuKEA	2022	Faster-RCNN	BERT	LXMERT	VQA2.0, A-OKVQA, KRVQA
ViLT	2021	Transformer	BERT	ViT	VQA2.0
ALBEF	2021	Transformer	BERT	Cross-Modal Transformer	VQA2.0

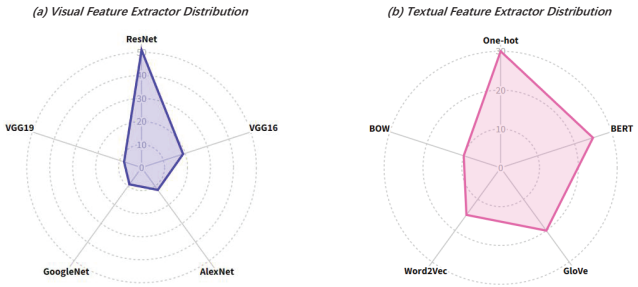


Fig. 3. Percentage distribution of the usage of visual and textual feature extractors.

with their corresponding size and position information into CNNs to extract region features. Compared with global-based features, this can better identify objects attribute, quantity, and location relationship [4]. The commonly used object recognition techniques is Faster RCNN [130].

For VQA tasks with video input, temporal channels are typical employed for visual feature extraction. Two common approaches are utilized for this purpose. One is C3D [155], which scales convolutional networks to three dimensions to capture temporal information effectively. Alternatively, optical flow [32] can be used to extract video features. This method analyzes pixel movement and frame-to-frame correlation to capture temporal dynamics. Please refer to Table 2 and Figure 3 for a summary of various methods for extracting visual features.

2.1.2 Textual Feature Extraction. Another crucial aspect is the extraction of texture features. Classic models for text encoding include **Bag-of-Words** (**BoW**) and Word2Vec [114]. With the emergence of **Recurrent Neural Networks** (**RNNs**), models like LSTM, Bi-LSTM, and GRU have been extensively utilized for handling sequential data. In addition, CNNs, originally prominent in computer vision, have also been applied for text feature extraction. Word2Vec, RNN, and CNN are all common models for extracting textual features until the advent of Bert [24]. Bert, namely Bi-directional Encoder Representation from Transformers, is a pre-training model proposed by Google AI in 2018. Bert's architecture is based on bidirectional Transformer components, and remarkably, fine-tuning with an additional output layer achieves outstanding performance across various downstream tasks. Bert stands as a milestone work in the history of NLP.

We give a summary of textual feature extraction models in Table 2, and conclude the visual and textual feature extractor distribution in Figure 3 to better show what the most used feature extraction methods are.

2.1.3 Improvement of Embedding. Numerous methods have been introduced for extracting embeddings of different modalities. To adapt to more complex questions and generate more accurate answers, significant improvements have been made in feature embedding, including Improving Visual Embedding [52, 53], Adding Additional Information [41, 85], and Improving Input Structure [67, 171]. We give the detailed improvement methods in Appendix A.

2.2 Fusion Mechanism

2.2.1 Simple Vector Operations. Traditional fusion methods typically utilize simple vector operations like element-wise addition, element-wise product, and vector concatenation to directly operate on visual features v_I and text features v_Q for generating joint representations. For element-wise addition and product, v_I and v_Q need to be related in the same dimension. In cases where they are not, a linear projection [115] can be employed to embed these vectors into the same space using transformation matrices W_v and W_q. Vector connection splices v_I and v_Q together as fusion vector v_F.

In general, the use of element-wise addition does not increase extra computation, and is, therefore, a common feature fusion method. In contrast, vector concatenation increases computational complexity significantly.

2.2.2 Deep Learning Method. The fusion of visual and textual features using deep learning can be categorized into three main approaches: non-attentional deep learning, attention-based deep learning, and GNNs.

Non-Attention Based Deep Learning. CNNs and RNNs are commonly utilized for non-attention-based multimodal fusion. Ren et al. [128] propose aligning image features v_I with word embeddings, feeding them into an LSTM model alongside question words v_{Q_i}. Ma et al. [107] leverage memory-augmented neural networks for VQA, using memory modules to maintain longer-term information.

CNN-based approaches have also been explored. Ma et al. [109] proposed an end-to-end framework incorporating image, sentence, and multimodal CNNs to enhance image-question interplay. Noh et al. [119] use a modified VGG-16 for image feature extraction and GRU for text, enabling adaptive parameter prediction. To preserve spatial information, Gao et al. [40] introduce Question-Guided Hybrid Convolution, using question-guided kernels to learn discriminative multimodal feature representations.

Bilinear pooling is widely employed for fine-grained visual recognition. Fukui et al. [35] introduce **Multimodal Compact Bilinear Pooling** (**MCB**) to compress bilinear models, efficiently combining multimodal features. Schwartz et al. [134] use MCB with attention mechanisms to capture high-order correlations, while Yu et al. [181] propose **Multimodal Factorization Bilinear** (**MFB**) pooling for effective fusion, employing sum pooling to reduce dimensionality.

Attention Based Deep Learning. Based on the information in the question, there are parts of the image that are more relevant to the question, which is also the part of the model that needs more attention. Therefore, the attention mechanism is introduced into the multimodal fusion mechanism of VQA, and Figure 4 compares the non-attention and attention-based deep learning methods. Li et al. [83] propose QRU, iteratively updating question representations based on relevant image regions. Shih et al. [141] introduce edge boxes to obtain image regions, and selectively combine image region features with text features, marking an early instance of attention-based deep

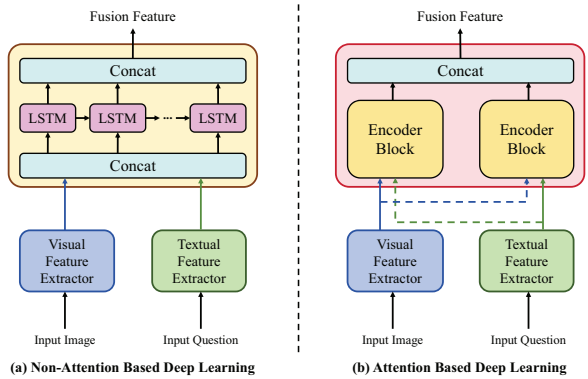


Fig. 4. Comparison of fusion mechanism utilizing (a) non-attention based deep learning and (b) attention based deep learning.

learning methods in VQA. It maps the region image features $V = (v_1, v_2, \ldots, v_m)$ and text features q to a common n-dimensional space, then calculates the inner product between regions and question answers to determine relative weights.

Researches establish loose, global associations between questions and images. Zhu et al. [193] propose LSTM-Att to establish semantic associations between image regions and text descriptions, capturing specific associations between image-related questions and regions. Attention models have evolved to capture finer-grained associations. However, it is not enough to only focus on local regions in visual features, and it is equally important to determine which words need to be focused on in the problem. Lu et al. [99] introduce co-attention mechanisms to jointly perform visual and question-guided attention. The joint attention is calculated as follows:

$$C = tanh(Q^T W_b V), \tag{1}$$

$$H^v = tanh(W_v V + (W_q Q)C), \tag{2}$$

$$H^q = tanh(W_q Q + (W_v V)C^T), \tag{3}$$

where V is a visual feature, Q is a text feature, C is co-attention, and W is the weight parameter.

Previous approaches mainly exploit low-level features while ignoring the rich semantics contained in high-level features. Yu et al. [181] present a multi-level attention network to reduce the semantic gap and visual attention for fine-grained spatial reasoning, aligning image regions and questions through feature multiplication for fine-grained spatial reasoning. To further enhance multimodal fusion, Nguyen et al. [118] propose a symmetrical co-attention mechanism, where each word in the question attends to image regions and vice versa. Wu et al. [162] addressed complex reasoning tasks by introducing a chain of reasoning model, enabling dynamic relational reasoning between objects.

GNN-Based Fusion Approaches. Traditional VQA methods often ignore the structural information in images or questions. Recent approaches leverage GNNs for better feature fusion. Norcliffe et al. [120] construct scene graphs conditioned on the question, using a K-kernel **Graph Convolutional Network (GCN)** to capture object interactions. Li et al. [80] propose **Relation-Aware Graph Attention Networks (ReGAT)**, encoding semantic, spatial, and implicit relations through attention mechanisms sensitive to node features and positional similarity:

$$\alpha_{ij} = \frac{\alpha_{ij}^b \exp((\mathbf{U}v_i')^T \mathbf{V} v_j')}{\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^b \exp((\mathbf{U}v_i')^T \mathbf{V} v_j')}. \tag{4}$$

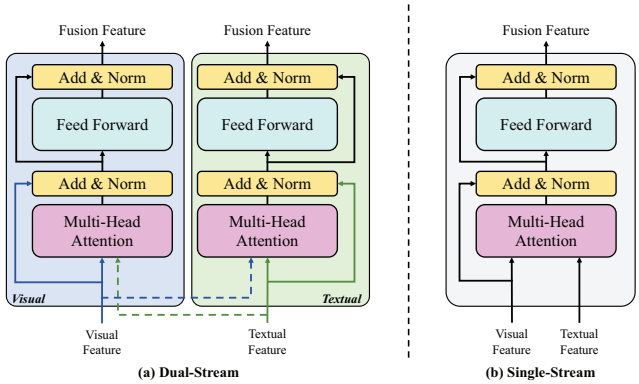


Fig. 5. Comparison of fusion mechanism utilizing (a) dual-stream pre-trained models and (b) single-stream pre-trained models.

For explicit relations, the attention mechanism accounts for edge labels and directions, allowing the model to capture richer semantic dependencies.

In addition to image graphs, Teney et al. [153] construct a question graph based on the token syntactic relations. Subsequently, a GRU-based GNN is deployed to aggregate first-order information, and a cross-modal attention mechanism aligns textual tokens $\{x_i'^Q\}_{i=1}^{N^Q}$ from the question graph with visual objects $\{x_i'^S\}_{i=1}^{N^S}$ of the scene graph as follows:

$$\alpha_{ij} = \sigma\left(W_5\left(\frac{x_i'^Q}{||x_i'^Q||} \circ \frac{x_i'^S}{||x_i'^S||}\right) + b_5\right). \tag{5}$$

Instead of using fully-connected graphs to represent images and questions, Huang et al. [51] prune edges in the visual and question graphs based on the object overlapping region and syntactic dependencies respectively. In the aggregation stage, a dual-channel GCN simultaneously captures relations between objects in images and relations among textual tokens in questions.

Transforming questions into instructions to guide scene graph learning has garnered recent attention. Shi et al. [140] propose using NLP tools to parse a given problem into a series of programs, and select different neural modules to infer the scene graph according to the corresponding programs. Since the instructions represented by the program set are finite and discontinuous, Hu et al. [50] parse questions into several textual instruction embeddings $\{c_t\}_{t=1}^{T}$ which are used to guide the process of message passing. In the procedure of scene graph learning, a GCN conditioned on instruction embeddings dynamically predicts edge weights $w_{j,i}^{(t)}$ to focus on different connections and aggregates information from neighboring nodes $\tilde{x}_{j,t}$ to the target $\tilde{x}_{i,t}$ in each iteration. Liang et al. [90] regard VQA as an answer generation task and propose a model LRTA consisting of four stages (Look, Read, Think, Answer). LRTA parses the problem into instructions using a transformer-based framework and traverses the scene graph using a recursive neural symbolic execution module that executes one instruction at each inference step.

2.2.3 Vision-Language Pre-Training Models. Vision-Language Pre-training models are trained on large-scale unlabeled data via self-supervision and fine-tuned for specific tasks, allowing for knowledge transfer and improved performance with minimal labeled data. In multimodal research, methods fall into three categories: Dual-Stream, Single-Stream, and other variants. Figure 5 provides a comparison of Dual-Stream and Single-Stream.

Dual-Stream. The dual-stream paradigm in multimodal research involves processing visual and textual inputs separately before integrating them through a cross-modal fusion module. Exemplifying this approach are ViLBERT [97] and LXMERT [151]. ViLBERT extends the BERT architecture to a multimodal cross-stream model, employing separate Transformers for visual and textual inputs and integrating them through a co-attention module. Similarly, LXMERT introduces an architecture to learn the language-vision connection, utilizing object relationship encoders and pre-training tasks to capture intra-modal and inter-modal relations.

In recent advancements, ERNIE-ViL [179] proposes integrating structured knowledge from scene graphs to enhance semantic alignment. This model leverages the structured knowledge obtained from scene graphs to facilitate fine-grained semantic understanding. Building upon previous research, Dou et al. [30] present the METER model, which refines the dual-stream approach by stacking transformer layers with self-attention, co-attention, and feedforward networks. METER conducts comprehensive experiments across various aspects of general pre-training models, providing insights into the efficacy of different architectural components.

Single-Stream. Different from Dual-Stream approaches, Single-Stream models integrate textual and visual inputs for semantic learning. Li et al. [74] propose Unicoder-VL, which matches specific text phrases with image features and inputs them jointly into a multi-layer Transformer for cross-modal representation learning. They introduce a formulation to combine text and image features, leveraging both region and location information. To mitigate overfitting with limited target tasks, Su et al. [148] train VL-BERT on large-scale image description and plain text datasets simultaneously, enabling the learning of more general feature representations. Their model employs stacked multi-layer Transformer modules to adaptively aggregate information from different modalities.

Chen et al. [17] propose UNITER, a general image-text representation learning model that conducts fine-grained semantic alignment between words and image regions. They introduce a novel pre-training task with conditional masking, enhancing the alignment process. Meanwhile, Kim et al. [64] present ViLT, a lightweight model focused on modal interactions without relying on region or deep convolutional features. Instead, they use a pre-trained ViT to extract visual features. In Single-Stream models, it's impractical to encapsulate intra-modal and vision-language learning in the same Transformer. Xue et al. [168] introduce self-attention to the visual domain to facilitate learning visual modalities. They utilize the Swin Transformer [96] to obtain visual feature embeddings and perform visual masking based on internal attention scores from the inter-modal Transformer.

Previous pre-trained models rely heavily on the image feature extraction process (such as Faster R-CNN [130] and ResNet [47]), requiring high hardware equipment and time-consuming training. Kim et al. [64] propose the lightweight model ViLT, the main computation of which is concentrated on the modal interactions. ViLT does not use region features or deep convolutional features but uses a pre-trained ViT model [29] to extract visual features.

$$\hat{z}^d = MSA(LN(z^{d-1})) + z^{d-1}, \tag{6}$$

$$z^d = MLP(LN(\hat{z}^d)) + \hat{z}^d, \tag{7}$$

where z indicates the vision-language embedding, LN means LayerNorm, and MSA indicates multiheaded self-attention.

Variants and Improvements in Vision-Language Pre-Training Approaches. There are some variants and improvements in training techniques, with the detailed introduction in Appendix B.

Early methods such as ViLBERT [98] employ task-specific networks for vision-language multi-task learning. Lu et al. introduced a shared backbone with task-specific layers for each

task, optimizing model parameters across different vision-language tasks.To enhance model robustness, adversarial learning was applied to vision-language pre-training by Gan et al. [36], who proposed the VILLA framework. This method introduces adversarial noise into the image and word embedding spaces, improving model generalization. Contrastive learning is introduced in the multimodal domain by Li et al. [84]. UNIMO leverages a three-stream architecture to process language, visual, and cross-modal fusion independently, optimizing representations for both single- and multi-modal tasks. The objective function involves maximizing the similarity between positive pairs of image and text embeddings:

$$\mathbb{E}_{V,W}\left[-\log \frac{\exp(d(V^+, W^+)/\tau)}{\sum_{\mathcal{X}} \exp(d(V', W')/\tau)}\right]. \tag{8}$$

Radford et al. [126] introduced CLIP, a significant step forward in learning image-text relationships through contrastive learning. CLIP normalizes word and region embeddings and computes their similarity via dot product, learning robust representations for zero-shot vision-language tasks, where I_e and T_e represent the joint multimodal features:

$$I_e = L2Norm(I_f, W_i), \quad T_e = L2Norm(T_f, W_t). \tag{9}$$

Li et al. [79] further improved vision-language models with ALBEF, which leverages momentum distillation to align image-text pairs and combat noisy data, introducing image-text contrastive loss. Recent models like BLIP [78] and PNP-VQA [154] have further extended this line of work by integrating multimodal generation and understanding within a unified framework, demonstrating strong performance in zero-shot VQA tasks.

These vision-language pre-training methods have not only bridged the gap between vision and language understanding in VQA but have also become essential for aligning visual and textual information. This alignment serves as a foundational element for transitioning from purely text-based Large Language Models to multimodal Large Language Models.

3 Comprehension of Image and Text with LLMs in Zero-Shot VQA

As text **large language models** (**LLMs**) have shown amazing performance in multiple textual tasks and attracted great attention from the whole community, more and more research attempts have been made to explore the introduction of LLMS into other domains [3, 88]. However, LLMs cannot directly process image information, so there is a rising need for generalized multimodal large language models (MLLMs) to accomplish various multimodal tasks [65, 189].

3.1 LLM Aided Visual Understanding

Since Large Language Models (LLMs) are text-based, early Multimodal Large Language Models (MLLMs) explore leveraging LLMs as central controllers for multimodal tasks [45]. LLMs act as a central controller that (1) analyze the prompt and history of dialogues, (2) split a complex task into simpler sub-tasks and (3) assign these tasks to appropriate models. Figure 6 shows the LLM aided visual understanding models. For instance, Microsoft Visual ChatGPT [163] integrates LLMs with various **Visual Foundation Models** (**VFMs**) (e.g., BLIP [78], Stable Diffusion, ControlNet), using a Prompt Manager to manage input/output formats. The LLM analyzes prompts, divides tasks, and invokes VFMs to generate outputs. Additionally, Visual ChatGPT utilizes dialogue history management and iterative reasoning to invoke further VFMs for more accurate results.

MM-REACT [175] focuses on broader visual interpretation by incorporating Azure APIs for tasks like celebrity recognition and Bing search. This enhances the LLM's role as a controller for visual understanding and interaction. As tasks grow more complex, LLMs evolve into decision-makers, as in IdealGPT [178], where autonomous agents handle complex tasks. These

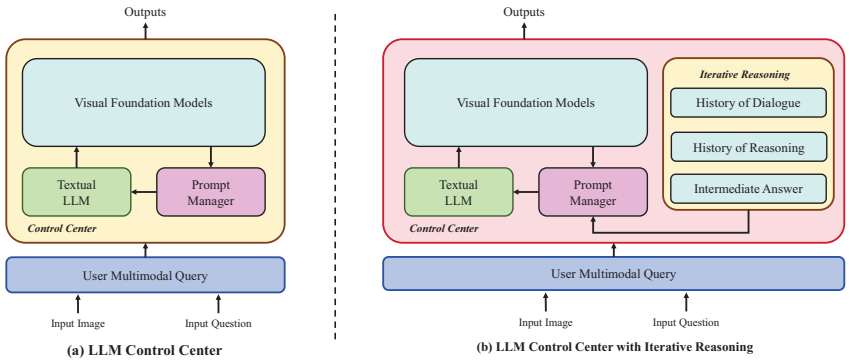


Fig. 6. Two architectures of the LLM aided visual understanding models.

agents consist of modules for profiling, memory, planning, and action [138], allowing the LLM to understand dynamic environments and organize responses effectively.

3.2 Image-to-Text Generation for Visual Understanding

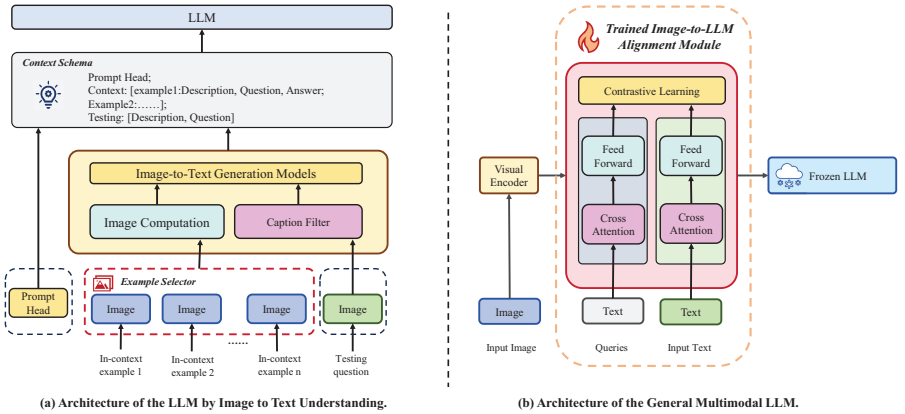
For the issue that textual large language models cannot process images, another inspiration is to transform images directly into corresponding textual descriptions [172], as shown in Figure 7. This typically involves models like Image Captioning and **Optical Character Recognition** (**OCR**) [147]. For example, PICA [172] uses image captioning to generate textual descriptions, which are concatenated with questions and fed into the LLM for question answering. To improve In-context Learning, PICA selects 16 training examples closest to the current test image-question pair using CLIP [126].

However, converting images into text can lead to inaccuracies or loss of essential visual details. To address this, IMG2LLM [44] generates more relevant captions and question-answer examples directly from the image. This model focuses on selecting image regions pertinent to the question and refining the captions for accuracy. It also synthesizes question-answer pairs to provide more representative in-context prompts. Prophet [137] enhances the selection of in-context examples and the generation of answer heuristics for VQA tasks. Prophet uses a Vanilla VQA [58] model to generate answer candidates, which serve as examples. These answer-aware heuristics, along with the testing samples, are input into the LLM as prompts, improving VQA performance.

3.3 General Multimodal LLM

LLM-aided visual understanding and image-to-text generation methods demonstrate the ability to leverage pure-text LLMs for multimodal tasks. However, these approaches do not inherently grant LLMs image comprehension capabilities. Instead, LLMs either rely on VFMs for image processing or convert images into textual representations. To address this, there is growing interest in developing general-purpose MLLMs with direct image comprehension abilities by integrating additional modalities into a unified framework, as shown in Figure 7.

3.3.1 The Training Paradigm of General Multimodal LLM. The training architecture for general multimodal LLMs, typically using frozen LLMs, follow a two-stage process, as shown in Figure 7. First, the LLM is frozen while external components, such as visual encoders and alignment modules, are trained to map image features into the textual space of the LLM [160]. In the second stage, the visual components are frozen, and the LLM is fine-tuned with multimodal data, often using techniques like LoRA or Q-LoRA [23]. This method significantly reduces the cost of extending LLMs to multimodal tasks, with a focus on designing effective image-to-text alignment modules,



(a) Architecture of the LLM by Image to Text Understanding.

(b) Architecture of the General Multimodal LLM.

Fig. 7. Architecture of the LLM by image to text understanding and general MLLM.

instead of training a new multimodal large model. Thus, it plays a pivotal role in the advancement of MLLMs, with the core challenge of designing more effective Image-to-Text alignment modules.

3.3.2 Image-to-Text Alignment in Multimodal LLM. Flamingo [3] introduces a vision encoder and a Perceiver Resampler to generate a fixed-length feature sequence, integrated with cross-attention layers for improved visual-textual alignment. PaLM-E [31] integrates visual features into the pre-trained PaLM model, leading to robust performance across real-world tasks, a method adopted by models like LLaVA [93] and Shikra [15]. However, freezing the LLM during training can limit alignment effectiveness. BLIP-2 [77] addresses this by proposing the Q-Former, using a two-stage pre-training approach where the visual encoder learns critical visual features through contrastive learning and image-text matching. This method enhances zero-shot capabilities, though it shows limited in-context learning improvements.

To reduce computational demands, smaller and faster models like PaLI-3 [16], based on SigLIP [184], offer competitive performance across multimodal benchmarks while requiring fewer resources.

3.3.3 Advanced Closed and Open-Source General MLLMs. As MLLM continues to develop, researchers continue to expand the data and parameter scale for MLLM pre-training, and continue pre-training, hoping to teach MLLM more and more new general knowledge and make it more powerful. In this context, the development of MLLM is no longer based on the improvement of VQA, a downstream task, but more focused on its general capabilities and performance in multiple tasks in the entire multimodal field. Many commercial models have emerged, such as OpenAI's GPT series, including GPT-4v with image and text interleaving processing capabilities [18], and GPT-4o frameworks with more powerful multimodal general capabilities and support for more multimodal inputs [56], such as Google's Genimi and Gemini Pro series [152].

In parallel, the open-source community has developed several high-performing models that rival, and in some cases surpass, these commercial counterparts. The **mPLUG series** [177], for example, has been praised for its versatile cross-modal capabilities. mPLUG-OWL excels in open-world vision-language tasks, benefiting from its ability to handle a wide range of image-text pairs and unstructured data sources. The **InternVL series**, such as InternVL-2 or InternLM-XComposer2.5-VL [19], distinguishes itself by focusing on improving vision-language alignment through sophisticated pre-training techniques, allowing for highly effective cross-modal understanding, which has proven especially effective in sophisticated tasks like math reasoning. Similarly, the **LLaVA series**

Table 3. MLLM Models with their Base LLM Models, MLLM Types, Techniques, and Performances

Models	Year	Base	Type			Technique			Setup	Performance on VQA	
			LLM-aid	img2txt	general mm	SFT	ICL	CoT		OK-VQA	VQAv2
Viusal ChatGPT [163]	2023	ChatGPT text-davinci-003	✓			✓		✓	-	Case study	
MM-REACT [175]	2023	gpt-3.5-turbo	✓			✓		✓	-	Case study	
PICa [172]	2022	GPT-3 (175B)		✓			✓		few-shot	46.9	-
		GPT-3 (175B)		✓			✓		few-shot	48.0	-
Img2LLM [44]	2023	OPT-3 (66B)		✓			✓		zero-shot	43.2	60.3
		OPT-3 (175B)		✓			✓		zero-shot	45.6	61.9
Prophet [137]	2023	GPT-3 API		✓			✓		few-shot	61.1	-
Flamingo [3]	2022	Chinchilla(70B)			✓		✓		zero-shot	50.6	56.3
		Chinchilla(71B)			✓		✓		few-shot(4)	57.4	63.1
		Chinchilla(72B)			✓		✓		few-shot(32)	57.8	67.6
BLIP-2 [77]	2023	OPT(6.7B)			✓	✓			zero-shot	36.4	52.6
		FlanT5(XXL)			✓	✓			zero-shot	45.9	65.0
LLaVA-1.5 [93]	2023	Vicuna(7B)			✓	✓			few-shot	-	78.5
Qwen-VL-Chat [7]	2023	Qwen(7B)			✓	✓			few-shot	56.6	78.2
mOLUG-owl2 [177]	2023	LLaMA-2(7B)			✓	✓			few-shot	57.7	79.4
InternVL2 [19]	2024	InternLM2-Chat			✓	✓			few-shot	-	-

[93] has garnered attention for its innovative approach to integrating large-scale vision-language models. Through a carefully curated combination of instruction tuning and multimodal dialogue datasets, the LLaVa series has achieved strong performance in both visual comprehension and interactive reasoning.

3.4 Techniques of the MLLMs

We summarize the recent MLLM models with their base LLM models, MLLM types, Techniques and Performances in Table 3. There are various techniques that have helped the researchers in generalizing LLM to MLLM, which equip the LLM with the image understanding capability.

Fine-Tuning and Instruction-Tuning. Fine-tuning enhances MLLM's ability to interpret input images and questions [57, 133, 161]. Typically, a VQA prompt includes: (1) task background description, (2) input context (image, image description, OCR), (3) the target question and answer candidates, and (4) reference answers [14, 44, 192]. The prompt design varies across models. For instance, Visual ChatGPT [163] incorporates multiple visual model formats, while domain-specific VQA may include specialized background knowledge or pre-labeled visual prompts for object detection and segmentation tasks [169].

Fine-tuning with specific datasets helps MLLM perform specialized tasks. For example, general MLLMs struggle with mathematical reasoning [100], so tailored datasets for diagram understanding [188], geometry reasoning [89], and handwritten equation recognition [124] have been created to improve performance. To avoid losing general capabilities, fine-tuning often involves both task-specific and general datasets.

In-Context Learning. To improve few-shot or zero-shot learning in VQA, researchers leverage **in-context learning** (**ICL**), which uses example contexts to enhance LLM performance [27, 45, 138]. Early ICL prompts were textual, selecting image-question pairs similar to the target [30, 103, 137], while more recent approaches generate examples directly from the target image itself [44]. Multimodal ICL integrates additional visual context, beyond text, to aid learning [106, 191]. Current research explores two key aspects: (1) improving example selection for better task understanding, with methods like RAG utilizing retrieval and generation to create more representative examples [137, 156, 172]; and (2) designing richer multimodal context schemas, such as MMICL, which generates subgraphs of target images with symbolic correspondences to textual elements [103, 191].

Visual Perception Capability. Early studies directly extend LLM to equip with visual functions and choose to call basic visual models, such as Object Detection, Image Captioning, and OCR, to

Table 4. Knowledge Reasoning Models

Models	Knowledge Source	Knowledge Reasoning	
		Methods	description
Explicit Knowledge-based [158]	DBpedia	Conventional	SPARQL queries
FVQA [159]	WebChild, ConceptNet, DBpedia	Conventional	KB query triplet
Ask me anything [165]	DBpedia	One-hop	SPARQL queries
Out of the box [116]	ConceptNet, DBpedia	One&Multi-hop	fact scoring and source scoring, Graph-based
Passage Retrieval [125]	Wikipedia passage	One-hop	Dense Passage Retrieval
Transform-Retrieve-Generate [37]	Wikipedia passage	One-hop	Dense Passage Retrieval, generative model
Incorporating external knowledge [75]	ConceptNet	Multi-hop	Memory-based: dynamic memory network
visual knowledge memory [149]	Visual Genome	Multi-hop	Memory-based: key-value structural memory
Inner Knowledge-Based Img2Doc [81]	Inner Knowledge	Multi-hop	Memory-based
Mucko [194]	Inner Knowledge	Multi-hop	Graph-based: factual,visual,semantic graph
context-aware knowledge aggr. [76]	Wikipedia	Multi-hop	Graph-based: semantic graph
See is Knowing [127]	Others	Multi-hop	Implicit Reasoning: ERMLP mode
MuKEA [26]	Inner Knowledge VQA2.0	Multi-hop	Implicit Reasoning: knowledge graph completion

provide usable visual information for plain text LLM [137, 172]. Furthermore, researchers explored directly giving LLM visual understanding capabilities. By designing a visual encoder adapted to LLM, the ability to match images and texts is pre-trained in a large amount of image and text data [21, 73, 77]. The general MLLM achieves pixel-level image understanding based on image encoders such as ViT. Compared with the previous MLLM based on object-level image understanding, it has better image-text alignment and fine-grained understanding capabilities [93]. Based on this training, MLLM enhances the semantic segmentation capability of images through better visual annotation [163], and combines multiple and larger image data to train MLLM to understand more diverse images [177]. However, since a large amount of training data comes from natural images in real scenes, MLLM exposes its defects in understanding specific images, such as understanding document images such as PDF (especially formulas in images) [25], understanding related text and symbols in real scenes [87], and understanding complex charts [183]. More real-scene OCR technologies are being explored.

4 Knowledge Reasoning in VQA

4.1 Knowledge Sources

External knowledge is indispensable for knowledge-based VQA task whose answers cannot be readily inferred from images but requires common sense knowledge. We summarize the most widely used knowledge sources for recent knowledge-based VQA methods, which can be divided into two categories based on the data format, structured knowledge such as DBpedia [6] and ConceptNet [94], and unstructured knowledge [26, 82]. We give a detailed introduction of the knowledge sources in Appendix C., and conclude the different knowledge source in models in Table 4.

4.2 Knowledge Extraction

4.2.1 Entity-Based Extraction. Entity-based methods extract visual or textual concepts from image-question pairs, using them as anchors for knowledge extraction. For instance, Wu et al. [165] extract key attributes from images and generate SPARQL queries for DBpedia retrieval. Similarly, Wang et al. [158] detect visual concepts (objects, attributes, scenes) and link them to synonymous entities in DBpedia to construct RDF graphs. To enhance extraction, Su et al. [149] apply subgraph hashing to match knowledge triplets with question phrases and expand relevant connections. Recently, Li et al. [76] introduced an approach that treats visual and textual concepts as KG anchors, expanding to include first-order neighbors and scoring nodes to select relevant knowledge.

4.2.2 Feature-Based Extraction. Feature-based methods focus on converting knowledge into continuous representations. Narasimhan et al. [117] use LSTMs to extract key relations from questions and calculate affinity scores between the image-question features and knowledge base facts,

selecting the most relevant knowledge. Ziaeefard et al. [195] further develop this by transforming facts into semi-phrases, which are represented using BERT. Ding et al. [26] merge knowledge extraction with image-question feature extraction through a pre-trained visual-linguistic model. Li et al. [82] view knowledge extraction as graph representation learning, linking images to captions and KGs, and using Deepwalk to generate knowledge-aware embeddings. More details about the knowledge extraction methods can be found in Appendix D.

4.3 Knowledge Reasoning

Knowledge reasoning in knowledge-based VQA involves deriving answers from extracted knowledge [139].

4.3.1 Conventional Reasoning Methods. Conventional methods often employ rule-based or template-based approaches. Wang et al. [158] extract visual concepts from images and parse questions using templates to generate SPARQL queries for answer reasoning. Wang et al. [159] take this further by using LSTMs to parse questions into KB query triplets, filtering concepts and relations from both the image and the knowledge base, and applying distinct rules based on the source of the knowledge.

4.3.2 One-Hop Reasoning Methods. Methods based on conventional reasoning are usually unstable under complex conditions and unscalable to different domains. One-hop reasoning methods utilize deep learning to solve these problems from first-order knowledge, where latent rules are learned for fact selection in training.

As for structured knowledge, Wu et al. [165] proposed an encoder-decoder answer generation architecture. They first generate SPARQL queries based on an image for KB searching. Then, collected knowledge comment paragraphs are combined together and sent to Doc2Vec [69] to get the knowledge feature. Finally, knowledge and image features are combined with question tokens, which are sequentially fed into an encoder-decoder LSTM framework to get the answer. Narasimhan et al. [116] treat answer reasoning as the combination of fact scoring and source scoring. They first extract the image feature, visual caption feature, and question feature, and utilize a **Multi-Layer Perception** (**MLP**) to fuse these features into an image-question representation $g^{MLP}_{w_1}(x, Q)$. Then, they leverage $g^{MLP}_{w_1}(x, Q)$ to score vectorized external knowledge $g^F(f_i)$, and choose the most relevant fact as the candidate answer $\hat{f}$:

$$\begin{aligned} S_{w_1}(g^F(f_i), g^{MLP}_{w_1}(x, Q)) &= cos(g^F(f_i), g^{MLP}_{w_1}(x, Q)) \\ &= \frac{g^F(f_i) \cdot g^{MLP}_{w_1}(x, Q)}{\|g^F(f_i)\| \cdot \|g^{MLP}_{w_1}(x, Q))\|}. \end{aligned} \tag{10}$$

Finally, the problem Q is sent into an LSTM to predict the answer source $\hat{s} = h^s_{w_2}(Q)$, where $\hat{s} \in \{Image, KB\}$. If $\hat{s} = Image$, the head entity of $\hat{f}$ is taken as the answer, and vice versa.

Qu et al. [125] take the idea of **Dense Passage Retrieval** (**DPR**) [63] and propose a coarse-grained approach for outside-knowledge-based VQA, that is, finding several passages containing answers as output. They use a BERT to extract the passage representations of candidate collection before the training process and store them in the memory slot to reduce redundant computations. To further merge the visual problem information, the given image and question are simultaneously fed into a pre-trained visual language model LXMERT [151] to obtain a query representation. The dot products of query representation and passage representations are calculated to obtain the top-k passages.

Since coarse-grained passages only provide texts that may contain an answer but not the answer itself, Gao et al. [37] take candidate passages as their external knowledge, and deploy a

generative model for answer reasoning. To enhance multimodal compatibility, they first utilize the combination of caption text C_i, attribute text L_i, and OCR text O_i to represent the given image $v_i = (C_i, L_i, O_i)$. Then, each candidate passages $p_{i,k}$ of top-k collections are concatenated with question context Q_i and image context v_i to a Transformer-based encoder to get k hidden layer representations:

$$\mathbf{z}^{Q_i} = \left(\mathbf{z}_1^{Q_i}, \mathbf{z}_2^{Q_i}, \cdots, \mathbf{z}_k^{Q_i}\right), \tag{11}$$

$$\mathbf{z}_k^{Q_i} = E_{SelfAttn}(Q_i, v_i, p_{i,k}). \tag{12}$$

Finally, these representations are sent to a decoder to generate the answer, and an auto-regressive cross-entropy loss is used to train the entire model.

$$P(a_1), \ldots, P(a_1) = \sigma\left(D_{SelfAttn}(\mathbf{z}^{Q_i})\right). \tag{13}$$

4.3.3 Multi-Hop Reasoning Methods. One-hop reasoning methods typically extract shallow knowledge from the KB and are incapable of exploiting implicit knowledge and handling inter-fact relations. To solve these problems, multi-hop reasoning has been widely used in recent methods, which refers to performing multi-step inference on the KB to explore the logical and semantic relations between facts. There are three main branches of multi-hop reasoning: memory-based reasoning, graph-based reasoning, and implicit reasoning.

Memory-Based Reasoning. Memory-based methods treat reasoning as the process of knowledge memory. These methods iteratively memorize candidate facts to extract relevant knowledge and ignore extraneous knowledge, which brings the capacity of multi-hop reasoning.

Li et al. [75] first filter the extracted knowledge, where facts are scored based on the knowledge graph topology and the top-N facts are selected as candidates. Then, the representations of candidate facts are extracted and stored in memory slots for reading and writing. In the next process, they deploy a dynamic memory network with T iterations for knowledge accumulation, which consists of an attention component and a memory updating component. The attention component assigns weights for each knowledge representation M_i:

$$\begin{aligned} \alpha^{(t)} &= softmax\left(\mathbf{w} tanh\left(\mathbf{W}_2 \mathbf{z}_i^{(t)} + \mathbf{b}_2\right)\right), \\ \mathbf{z}_i^{(t)} &= \left[M_i; m^{(t-1)}; \mathbf{q}\right], \\ \mathbf{q} &= tanh\left(\mathbf{W}_1\left[\mathbf{f}^{(I)}; \mathbf{f}^{(Q)}; \mathbf{f}^{(A)}\right] + \mathbf{b}_1\right), \end{aligned} \tag{14}$$

in which $\mathbf{f}^{(I)}, \mathbf{f}^{(Q)}, \mathbf{f}^{(A)}$ are features of images, questions, and multi-choice answers respectively. The memory updating component accumulates knowledge based on attention weights to update the memory vector $m^{(t)}$:

$$\begin{aligned} m^{(t)} &= RELU\left(\mathbf{W}_3\left[m^{(t-1)}; \mathbf{c}^{(t)}; \mathbf{q}\right] + \mathbf{b}_3\right), \\ \mathbf{c}^{(t)} &= \sum_{i=1}^{N} \alpha^{(t)} M_i, t = 1, \ldots, T. \end{aligned} \tag{15}$$

Finally, $\mathbf{f}^{(I)}, \mathbf{f}^{(Q)}, \mathbf{f}^{(A)}$ and $m^{(T)}$ are fused together to obtain the confidence score for each candidate answer.

Instead of directly considering each fact as an entirety, Su et al. [149] use key-value structural memory slots. They first decompose each knowledge triplet (s, r, t) into three key-value pairs (i.e.,

(s, r)-t, (s, t)-r, (r, t)-s), which are passed to the joint embedding module to get the key representation $\mathbf{k}_i$ and value representation $\mathbf{v}_i$:

$$\mathbf{k}_i = \Psi(e_1, u_i) + \Psi(e_2, u_i), \ \mathbf{v}_i = \Psi(e_3, u_i), \tag{16}$$

where e_1, e_2, e_3 are different entries of i-th triplet, and u_i is the image feature. Then, a memory network iteratively refines the memory vector by performing key addressing and value addressing to obtain the answer. Li et al. [81] store all entries of each knowledge triplet in a value slot (i.e., $[F_s, F_r, F_t]$) and take their average representation as the key embedding. Then, a memory reading module captures the correlation between query embedding $\hat{q}$ and key-value pairs ($\hat{k}_i$-$\hat{v}_i$) and obtains the question-aware knowledge representation m^t which is further used to guide the graph learning:

$$\begin{aligned} m^t &= \sum_{i=1}^{N} p_i \hat{v}_i, \\ p_i &= softmax\left(\hat{q} \cdot \hat{k}_i^T\right), \\ \hat{v}_i &= \sum_{\hat{v}_{ij} \in [F_s, F_r, F_t]} \left(1 - softmax(\hat{q} \cdot \hat{v}_{ij}^T)\right) \hat{v}_{ij}/2. \end{aligned} \tag{17}$$

Graph-Based Reasoning. Graph-based methods tend to represent extracted facts as graphs, and then use the message passing paradigm to aggregate knowledge from multi-hop neighborhoods to the target nodes. This process enables the model to explicitly exploit the attribute information and relational information embedded in the KB, which is the most commonly used multi-hop reasoning method for the knowledge-based VQA task.

Narasimhan et al. [116] propose a GCN based model for graph-based reasoning, which mainly consists of two components: the factual graph construction and answer scoring. For the factual graph construction, facts fetched from the KB are composed into a homogeneous graph, where each node represents an entity, and each edge represents a relation. For incorporating images and questions information, the representation of each node is formed by the concatenation of image feature $g_w^V(I)$, question feature $g_w^Q(Q)$ and entity feature $g_w^C(e)$:

$$H_i^0 = \left(g_w^V(I); g_w^Q(Q); g_w^C(e)\right), \ e_i \in E, \tag{18}$$

where E is the entity set. In the process of answer scoring, a GCN integrates node features based on graph topology and outputs the Lth layer features $\hat{g}(e_i) = H_i^L$ which are fed into an MLP to predict answer $\hat{A}$:

$$\hat{A} = \arg\max_{e_i \in E} MLP(\hat{g}(e_i)). \tag{19}$$

In addition to using graph structure to represent the structural relationships between facts, Zhu et al. [194] further introduce visual graphs and semantic graphs to comprehensively depict image information. The visual graph is a scene graph, which is composed of objects in the given image and their positional relations. As for the semantic graph, an image is first sent to the DenseCap [61] to generate several captions, which are parsed into a semantic graph. The knowledge reasoning of this model comprises two parts: intra-modal knowledge selection and inter-modal knowledge reasoning. In the first stage, the visual, semantic, and factual graphs are aggregated separately using different GCNs to obtain the updated node features: $\{\hat{v}_i^V\}_{i=1}^{N^V}$, $\{\hat{v}_i^S\}_{i=1}^{N^S}$, $\{\hat{v}_i^F\}_{i=1}^{N^F}$ respectively. In the second stage, the information on the visual and semantic graphs is mapped to the factual graph (i.e., visual-to-factual and semantic-to-factual). Taking the semantic-to-factual process as an

example:

$$
\begin{aligned}
m_i^{S\to F} &= \sum_{j\in N^V} \Upsilon_{ji}^{S\to F}\hat{v}_j^S, \\
\Upsilon_{ji}^{S\to F} &= softmax\left(w_c tanh\left(W_8\hat{v}_j^S + W_9\left[\hat{v}_i^F, q\right]\right)\right),
\end{aligned}
\tag{20}
$$

in which $\Upsilon_{ji}^{S\to F}$ is the attention vector between $\hat{v}_i^F$ and each node of semantic graph, and $m_i^{S\to F}$ is the weighted sum of nodes on semantic graph for $\hat{v}_i^F$. Finally, $m_i^{S\to F}, m_i^{V\to F}, \hat{v}_i^F$ are fed into an element-wise gate network to get the final representation $\tilde{v}_i^F$:

$$
\begin{aligned}
\tilde{v}_i^F &= W_{11}\left(gate_i \circ \left[m_i^{S\to F}; m_i^{V\to F}; \hat{v}_i^F\right]\right), \\
gate_i &= \sigma\left(W_{10}\left[m_i^{S\to F}; m_i^{V\to F}; \hat{v}_i^F\right]\right).
\end{aligned}
\tag{21}
$$

Li et al. [76] treat knowledge reasoning as a process of anchor entity feature learning, where anchor entities could continuously acquire knowledge from the KB and facilitate answer prediction. They first use an object detector and a natural language parsing tool to extract entities from images, questions and candidate answers as anchor entities. In the second process, the first-order neighbors of anchor entities are obtained from the KB to form a global knowledge graph, which is fed into the attention-based GNN for feature aggregation. Then, extracted knowledge is distilled into three auxiliary features $\vec{e}^{(ctx)}, \vec{u}, \vec{e}^{(ans)}$:

$$
\vec{e}^{(ctx)} = \sum_{e_i\in C} \alpha_i\vec{e}_i,\ \vec{u} = RELU(\beta\cdot W_3),\ \vec{e}^{(ans)} = \sum_{e_j\in\mathcal{A}} \beta_j\vec{e}_j, \tag{22}
$$

$$
\text{where } \alpha_i \propto exp(h^q\cdot\vec{e}_i),\ \beta_j \propto exp(\vec{e}^{(ctx)}\cdot\vec{e}_j), \tag{23}
$$

in which h^q is the query embedding, C is the anchor entity set extracted from the question and image, and $\mathcal{A}$ is the anchor entity set extracted from candidate answers. Finally, auxiliary features are fused with image feature $\tilde{v}_k$ and h^q:

$$
\vec{f} = BaseFusion(\{\tilde{v}_k\}; \{\vec{e}^{(ctx)}, \vec{e}^{(ans)}, \vec{u}\}\cup\{h_m^q\}). \tag{24}
$$

Implicit Reasoning. Different from memory-based or graph-based methods, implicit reasoning treats the multi-hop reasoning task as an entity feature space learning issue, which aims to map the head, relation, and tail entities into a common feature space such that they can establish some statistical associations.

Ramnath et al. [127] proposed the "See is Knowing" framework, where the visual information is represented as several knowledge vectors to facilitate answer prediction. They first deploy an ERMLP model [28] on the large-scale KG, which captures the intrinsic connections between entities by learning to identify whether given facts exist. ERMLP can implicitly perform multi-hop reasoning in learning and generate a dense embedding for each entity. Then, scene, object, and action concepts are detected in given images and passed to ERMLP to get their knowledge-aware embedding $e_i^j, j\in[1,m]$. Finally, these visual knowledge embeddings and the query embedding $A(q_i)$ are passed into an attention module:

$$
A(I_i) = \sum_{j=1}^{m} \alpha_I^j e_i^j,\ \alpha_I^j = \frac{exp(w_{\alpha_I}^T[A(q_i); e_i^j])}{\sum_{k=1}^{m} exp(w_{\alpha_I}^T[A(q_i); e_i^k])}, \tag{25}
$$

in which $A(I_i), A(q_i)$ are further used to query the answer from the KG.

Ding et al. [26] address the rigidity of knowledge graphs (KGs) in understanding complex scenes by introducing MuKEA, which directly extracts and accumulates multimodal knowledge from

VQA scenarios. They frame the learning process as a multimodal knowledge graph completion task, leveraging image-question pairs to extract knowledge triplets and training the model based on the TransE framework [11]. For head entity extraction, a pre-trained visual-language model, LXMERT [151], jointly encodes image and question embeddings. These embeddings are processed through a hard attention mechanism to capture correlations between objects and question tokens. For tail entity extraction, the answer representation serves as the tail embedding. This approach enables implicit reasoning, allowing the model to continuously acquire multimodal knowledge from VQA datasets while uncovering latent relationships.

4.4 Multimodal Reasoning with LLMs

Recent years have witnessed the remarkable progress of MLLMs, especially in their surprising zero/few-shot reasoning abilities. In varieties of multimodal reasoning tasks (e.g., VQA), MLLMs often demonstrate impressive effectiveness. Therefore, MLLMs and VQA task run towards each other at the same time, forging a new direction for VQA research. Specifically, MLLMs are capable to make great comprehension and integration of information from image and textual questions after pretraining on large-scale multimodal data, enabling them to reason carefully and generate appropriate answers. Furthermore, several strategies have recently emerged to enhance the reasoning capabilities of MLLMs and improve the VQA performance, such as multimodal instruction tuning [21], multimodal in-context learning [3], multimodal chain-of-thought [103], and LLM-aided visual reasoning [138].

4.4.1 Multimodal Instruction Tuning. Pretrained MLLMs often struggle with generalizing to novel tasks and aligning with users' intentions, resulting in incorrect and dissatisfactory responses. To address these limitations, a strategy known as **multimodal instruction tuning** (**M-IT**) [133] has been introduced. Instruction refers to the task description. Multimodal instruction tuning is a technique that involves finetuning pretrained MLLMs on a collection of multimodal instruction-following data. Tuning in this way, MLLMs are guided to understand and adapt to the task of interest, thus boosting their zero-shot reasoning capabilities and task-specific performance. The success of some notable frameworks on VQA (e.g., BLIP-2 [77]) validates the effectiveness of this idea.

4.4.2 Multimodal In-Context Learning. As the demand for customized MLLMs for specific VQA task continues to grow, finetuning them by instruction tuning proves to be resource-intensive and may diminish the model's generalization capabilities. Furthermore, state-of-the-art MLLMs like GPT-4V are primarily accessible only through API calls, with their parametric weights remaining proprietary and unavailable to the public. This scenario underscores the growing need for a new methodology, **multimodal in-context learning** (**M-ICL**), which allow learning from analogy [27] without requiring parametric updates. Specifically in M-ICL, MLLMs learn from a few examples noted as demonstration. The examples not only provide supplementary contextual knowledge for MLLMs but also exert flexible control over output, thereby solving complex and unseen tasks in a few-shot manner [175]. Researches in M-ICL [3, 173] are shown empirically to enhance the reasoning ability of MLLMs on VQA tasks.

4.4.3 Multimodal Chain-of-Thought. In recent studies, chain-of-thought has gained widespread usage in eliciting the multi-step reasoning abilities of LLMs. Specifically, CoT aims to enable LLMs to imitate the step-by-step thinking process of humans. It encourages the LLMs to generate not only the final answer but also the intermediate reasoning chains that lead to the answer by adding a prompt like "Let's think step by step". Subsequently, to extend CoT reasoning to multimodality, several works [49, 132] have been proposed to extend the unimodal

CoT to **Multimodal CoT** (**M-CoT**). Given the inputs in different modalities, Multimodal CoT decomposes multi-step problems into intermediate reasoning steps (rationale) and then infers the answer. The success of multiple researches [102, 190] validates the effectiveness of M-CoT in enhancing MLLMs reasoning ability in VQA tasks.

4.4.4 LLM-Aided Visual Reasoning. Building on the achievements of tool-augmented LLMs [122], researchers have explored the potential of invoking external tools and modular approaches for visual reasoning tasks like VQA. Specifically, vision foundation models [163] like image captioning and OCR are usually invoked for better visual information understanding. Invoking external tools [103] such as knowledge retrieval and web search engines help LLMs access real-time information and leverage domain-specific knowledge from external resources. Multiple researches [138, 170] indicate that effective utilization of external tools enables LLMs to accommodate various reasoning capabilities for accomplishing complex VQA tasks.

4.4.5 Math and Logical Reasoning. Multimodal large language models (MLLMs) are increasingly applied to mathematical and logical reasoning tasks, which require not only linguistic understanding but also structured reasoning, abstract thinking, and precise computation–challenges for traditional language models [100, 167]. To address these limitations, various approaches have been explored. A prominent method is symbolic manipulation, where MLLMs are trained to recognize and transform symbolic representations such as formulas and diagrams [86, 166]. By integrating visual features from images or written equations with text-based inputs, these models enable nuanced reasoning in spatial and symbolic contexts. Efforts to enhance visual reasoning for mathematical diagrams and logical charts [188] further expand their capabilities. Another key approach is chain-of-thought reasoning [59, 68], where models extract symbolic rules from data and apply them in a structured, step-by-step manner, enabling tasks such as proofs and multi-step logical arguments. Recent advancements also leverage multi-agent systems, integrating external symbolic solvers like Wolfram Alpha. In this setup, MLLMs collaborate with specialized math engines to perform complex calculations or proofs, significantly enhancing accuracy in higher-level reasoning tasks.

5 Datasets and Metrics

5.1 Datasets

In this part, we systematically summarize the most widely used VQA datasets that are divided into two categories: (1) datasets whose questions are typically based on common sense knowledge (Section 5.1.1). (2) datasets based on external knowledge (Section 5.1.2). The statistics of all datasets are listed in Table 5, and we conclude the seven most widely-used dataset in Figure 8. Some representative datasets are selected to be presented here, and the rest are shown in Appendix E.

5.1.1 Datasets without External Knowledge.

DAQUAR [110]. DAQUAR is the first proposed VQA challenge. It is built on the NYU-Depth v2 [143] dataset, which contains 1,449 images and 12,468 Q&A pairs. Its annotation generation methods include synthetic and human. The synthetic annotation uses eight predefined templates and the original annotation of NYU-Depth v2. Human annotations come from 5 in-house participants. Although the proposal of DAQUAR is important to VQA, it also has some problems. For instance, the magnitude of images is too small; the image quality is poor and disorganized; the dataset is unbalanced; there are too many single-choice questions.

VQAv1 [5]. VQAv1 is one of the most widely used datasets, which contains 204,721 real images from the COCO dataset (123,287 images for training and 81,434 images for testing). It covers

Table 5. Datasets for VQA and their Main Characteristics

Dataset	Knowledge based?	Published year	Image source	Images	Q&A pairs	Avg.Q/ Imgae	Avg.Q length	Avg.A length	Questions Generation
DAQUAR [110]	✗	2015	NYU-Depth V2	1,449	12,468	8.6	11.5	1.2	Human
COCO-QA [129]	✗	2015	COCO	117,684	117,684	1.0	8.6	1.0	Automatic
Visual Madlibs [180]	✗	2015	COCO	10,738	360,001	33.5	4.9	2.8	Human
FM-IQA [38]	✗	2015	COCO	158,392	316,193	2.0	7.4	3.8	Human
VQAv1 [5]	✗	2015	COCO	204,721	614,163	3.0	6.2	1.1	Human
Visual Genome [66]	✗	2016	COCO & YFCC100M	108,077	1,445,322	13.4	5.7	1.8	Human
Visual7W [66]	✗	2016	Visual Genome	47,300	327,939	6.9	6.9	2.0	Human
VQAv2 [42]	✗	2017	COCO	204,721	1,105,904	5.4	6.1	1.2	Human
CLEVR [60]	✗	2017	Synthetic	100,000	999,968	10.0	18.4	1.0	Synthetic
CLEVR-CoGenT-A [60]	✗	2017	Synthetic	100,000	999,951	10.0	-	-	Synthetic
CLEVR-CoGenT-B [60]	✗	2017	Synthetic	30,000	299,972	10.0	-	-	Synthetic
VQA-CPv1 [2]	✗	2018	COCO	205,000	370,000	1.8	-	-	Human
VQA-CPv2 [2]	✗	2018	COCO	219,000	658,000	3.0	-	-	Human
VizWiz	✗	2018	blind people by phone	72,205	72,205	1.0	-	-	Human
VQA-Rephrasings [135]	✗	2019	VQAv2	40,504	162,016	4.0	-	-	Human
GQA [54]	✗	2019	COCO & Flickr	113,018	22,669,678	200.6	-	-	Synthetic
DocVQA [113]	✗	2021	UCSF Library	12,767	50,000	3.9	9.5	2.4	Human
InfographicVQA [112]	✗	2021	Internet	5,485	30,035	5.5	11.5	1.6	Human
IconQA [104]	✗	2022	digital textbooks	96,817	107,439	1.1	8.4	-	Human
PDFVQA [25]	✗	2023	PubMed PDF Doc	25,147	130,700	5.2	-	-	Automatic
E-VQA [176]	✗	2023	News article	2,690	9,088	3.4	-	-	Automatic
KB-VQA [158]	✓	2015	COCO & ImageNet	700	2,402	3.4	6.8	2.0	Human
FVQA [159]	✓	2017	COCO	2,190	5,826	2.7	9.5	1.2	Human
R-VQA [101]	✓	2018	Visual Genome	335,000	4,335,966	13.4	-	-	Human
KVQA [136]	✓	2019	Wikidata	24,602	183,007	7.4	10.1	1.6	Human
OK-VQA [111]	✓	2019	COCO	14,031	14,055	1.0	8.1	1.3	Human
ViQuAE [70]	✓	2022	Wikidata	3,300	3,700	1.1	12.4	-	Automatic
KRVQA [13]	✓	2022	Visual Genome	32,910	157,201	4.8	11.7	-	Automatic
Lora [39]	✓	2023	food-and-kitchen	100,000	200,000	2	-	-	Automatic

Avg.Q/Image indicates how many questions a image meanly corresponds to, and Avg.Q length and Avg.A length denotes the avarage length of question and answer respectively.

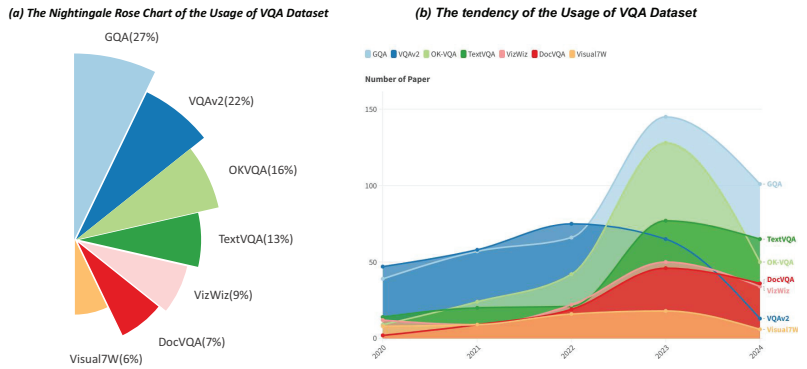


Fig. 8. The statistics of the widely-used dataset from 2020 to 2024.

614,163 free-form questions and 7,984,119 answers, allowing yes/no, multiple-choice, and open-ended forms of questions. These questions are collected by humans, and each question is manually annotated by 10 different people. The annotations also include the answers given by humans without looking at the images.

VQAv2 [42]. VQAv2 is the enhanced version of the VQAv1 dataset, which contains 204,721 images sourced from the COCO dataset. It has 443,757, 214,354, and 447,793 question annotations on the training set, validation set, and test set, respectively. VQAv2 has a total of 1,105,904 free-form Q&A pairs annotated by humans, twice as many as VQAv1, and provides a complementary image for each question so that the same question can be combined with two similar images to generate

different answers. Compared with VQAv1, VQAv2 reduces the bias and imbalance of the dataset through the above improvements.

CLEVR [60]. CLEVR is a synthetic dataset and contains 100,000 rendered images and about 1M synthetic Q&A pairs where 853,000 questions are totally different. To make the task more challenging, questions are divided into five categories: querying attributes, comparing attributes, existence, counting, and integer comparison and each image is represented as a visual scene composed of simple geometric bodies, where a VQA model needs to handle novel combinations of unseen attributes during training and goes through a long reasoning process to answer the question.

IconQA [104]. IconQA consists of 96,817 Icon images and 107,439 Q&A pairs. These Icon images come from IXL Math Learning, an open-source mathematics textbook on the Internet, and Q&A pairs are obtained by manual collection and filtering. The questions of this dataset are mainly divided into three subtasks: 57,672 multi-image-choice, 31,578 multi-text-choice, and 18,189 filling-in-the-blank. The questions of IconQA are derived from real-world mathematical questions, which require commonsense reasoning and arithmetic reasoning.

5.1.2 Datasets with External Knowledge Base.

KB-VQA [158]. KB-VQA is the first VQA dataset requiring an external KB, which includes 700 images from the COCO dataset and 2,402 Q&A pairs. KB-VQA has 23 templates for questions, and each question is proposed by five workers according to one of the appropriate templates. The proposers assign different labels to questions of different knowledge levels. Answering questions at the "KB-knowledge" level requires the use of a KB like DBpedia. The "KB-knowledge" level questions in KB-VQA are far more than that of other contemporaneous VQA datasets.

FVQA [159]. FVQA has 2,190 images and 5,826 questions which are split into five train/test sets (1,100/1,090 images and 2,927/2,899 questions for training/testing per set). The questions can be divided into 32 categories in total. Its annotations include not only Q&A pairs, but also extra knowledge. FVQA builds a KB by collecting knowledge triples from WebChild, ConceptNet, and DBpedia, which contains 193,449 sentences as supporting facts related to 580 visual concepts (234 objects, 205 scenes, and 141 attributes). This dataset contains a supporting fact in each Q&A pair.

OK-VQA [111]. The **Outside Knowledge-VQA** (**OK-VQA**) dataset consists of 14,055 questions (including 12,951 unique questions) and 14,031 real images from the COCO dataset. The labeling process of OK-VQA is divided into two steps: first, workers are asked to provide questions that require external knowledge to answer for a given image, and then five different workers are asked to label answers for each image-text pair. After the annotation is completed, further filtering is required. If the answer to a question have more than five Q&A instances, the question will be deleted, thereby ensuring an even distribution of answers and eliminating potential bias.

5.2 VQA Datasets with MLLM Benchmark

With the rapid advancement of multimodal large language models (MLLMs), increasingly sophisticated general-purpose benchmarks have emerged to evaluate their performance across various dimensions. These benchmarks are designed to assess a broad range of capabilities, often with a foundation in VQA annotation. The creation of such evaluation data typically involves filtering image data from extensive sources, followed by generating corresponding **question-and-answer** (**QA**) pairs, either through automated processes or manual annotation. Although these benchmarks are not always explicitly developed to evaluate methodologies specific to the VQA task, they can nonetheless be regarded as generalized VQA datasets due to their shared emphasis on image-based question answering.

One of the earliest unified MLLM benchmarks, MME [33], compiles a substantial collection of images and generates corresponding QA pairs to evaluate MLLM performance, emphasizing consistency and objectivity. Similar efforts have been observed in subsequent benchmarks, such as SEEDBENCH [72] and SEEDBENCH-2 [72], both of which aim to standardize MLLM evaluation. However, more recent benchmarks have shifted focus toward assessing specific capabilities, expanding the scope of evaluation to include more specialized dimensions. For example, recent benchmarks evaluate models on their ability to comprehend visual information [12, 34, 71], demonstrate reasoning skills [131, 187], and perform in-context learning [92, 142]. In addition, some benchmarks address challenges such as hallucination detection and mitigation [20, 92], where models are tested for their ability to provide accurate information without generating false or misleading content.

Moreover, recent developments in MLLM benchmarks have extended their evaluation scope beyond general comprehension tasks, venturing into highly specialized domains. These include mathematics, physics, music, medicine, and more, as seen in datasets such as MathVista [100], MMMU [182], and CMMMU [186]. These domain-specific benchmarks highlight the growing recognition that MLLMs must be versatile, capable of handling complex, domain-oriented tasks that require a deep understanding of specialized knowledge. This multidimensional evaluation framework ensures that MLLMs are not only assessed for their general performance but also their abilities to excel in high-stakes areas.

5.3 Evaluation Metrics

Common VQA evaluation metrics can be divided into two categories: objective evaluation metrics and subjective evaluation metrics, which are often used for two mainstream VQA tasks: open-ended and multiple-choice.

5.3.1 Subjective Evaluation Metrics. The most common human evaluation is to ask human judges to directly evaluate the quality of the generated answers, either from an overall perspective or from a specific dimension. These specific dimensions need to be able to reflect the interrelated properties of the answer sentences. For example, Bai et al. [8] evaluate from three dimensions: grammar, faithfulness, and coherence. If evaluating in terms of the entire answer, human judges are usually asked to select one of three levels of fine-grained evaluation as a score, including 0 (completely false), 1 (partially true), and 2 (exactly true).

On the FM-IQA dataset, Gao et al. [38] propose a **Visual Turing Test** (**VTT**) based on human evaluation. In this test, a lot of sets of images and questions, along with their corresponding human-annotated answers or answers generated by a VQA model, are presented to human judges. Human judges need to discriminate whether the answers are given by humans or computers based on the given materials. If an answer of the VQA system is judged to be that of a human, the answer passes the VTT. Finally, the percentage of answers that pass the VTT is counted, and several additional VTTs are set to calculate the standard deviation.

5.3.2 Objective Evaluation Metrics.

QA Evaluation Metrics. Simple Accuracy [196] based on string matching is first proposed. For the multiple-choice VQA task, the comparison between the predicted answer and the ground truth is straightforward. However, as generated answers are mostly phrases consisting of multiple words in the open-ended VQA task, Simple Accuracy is often difficult to evaluate. On the one hand, indiscriminate judgment is clearly flawed, because the severity of wrong answers varies. For example, compared with the completely irrelevant answers obtained from prediction, the severity of answers that contain temporal errors is significantly lower, where the punishment should be

different for these situations. On the other hand, there may be multiple matching answers to the same question, while their appropriateness may vary. For instance, the correct answer to "What's swimming in the water?" is "bluefish", while "scad" means the same as "bluefish", and "fish" is also appropriate.

Antol et al. [5] propose $Accuracy_{VQA}$, which is used for open-ended evaluation on the VQAv1 dataset:

$$Accuracy_{VQA} = \min\left(\frac{n}{3}, 1\right), \tag{26}$$

where n is the number of predicted answers that are the same as those given by annotators. This means that the answer predicted by the algorithm with the same answer as three or more annotations is 100% accurate. However, this evaluation standard is also unreasonable. For example, the annotators of the COCO-VQA dataset only have consensus on a few questions, which limits the highest accuracy of the model.

In addition, $Accuracy_{VQA}$ is prone to errors on "yes/no" type questions. In this type of question, the answer "yes" or "no" may repeat more than three times in one question, which would result in a high score of both "yes" and "no" on the $Accuracy_{VQA}$. Other Metrics are shown in Appendix F.

Generation Evaluation Metrics. Some of the metrics originally used to evaluate answer generation can also be used in the open-ended VQA task. The mainstream evaluation metrics among them typically include **Bilingual evaluation understudy** (**BLEU**) [121], **Recall-Oriented Understudy for Gisting Evaluation** (**ROUGE**) [91] and **Metric for Evaluation of Translation with Explicit Ordering** (**METEOR**) [9]. The effectiveness of using these generation metrics for VQA system evaluation has been confirmed in Refs. [1, 46]. We give a breif introduction of BLEU metric, and the ROUGE and METEOR are introduced in Appendix F.

BLEU measures the quality of the answer by comparing the coincidence degree of n-gram phrases of different lengths in the predicted answer and the real answer. The higher the coincidence degree is, the higher the quality of the answer is. BLEU calculates the coincidence accuracy of the corresponding answer according to the following formula:

$$CP_n(C, S) = \frac{\sum_i \sum_k min(h_k(c_i), max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k h_k(c_i)}, \tag{27}$$

where c_i is the candidate answer and its corresponding group of reference answers is $S_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\} \in S$, ω_k represents the possible n-grams of the kth group, $h_k(c_i)$ represents the number of occurrences of ω_k in the candidate answer c_i, and $h_k(s_{ij})$ represents the number of occurrences of ω_k in the reference answer s_{ij}. BLEU is the weighted geometric average of n-grams coincidence accuracy, which represents the ratio of the correct matching times of n-grams to the occurrence times of all n-grams. The calculation formula is as follows:

$$BLEU_N(C, S) = b(C, S) exp\left(\sum_{n=1}^{N} \omega_n log CP_n(C, S)\right), \tag{28}$$

where N=1,2,3,4, ω_k is generally $\frac{1}{n}$ for every n.

5.4 Examples and Comparisons of the Performance of Four Widely-Used Datasets

To show the unique contributions of different datasets and their roles within the evaluation framework, we analyze widely-used VQA datasets and MLLM benchmark. Figure 9 provides an overview of each dataset's key attributes alongside illustrative examples of image-related challenges. For each VQA dataset, we synthesize a comparative evaluation of model performance, including state-of-the-art approaches, as shown in Table 6. This analysis spans conventional deep learning architectures, attention-driven methods, vision-language pre-trained models, and MLLMs. A notable consideration in these comparisons is the variability in experimental

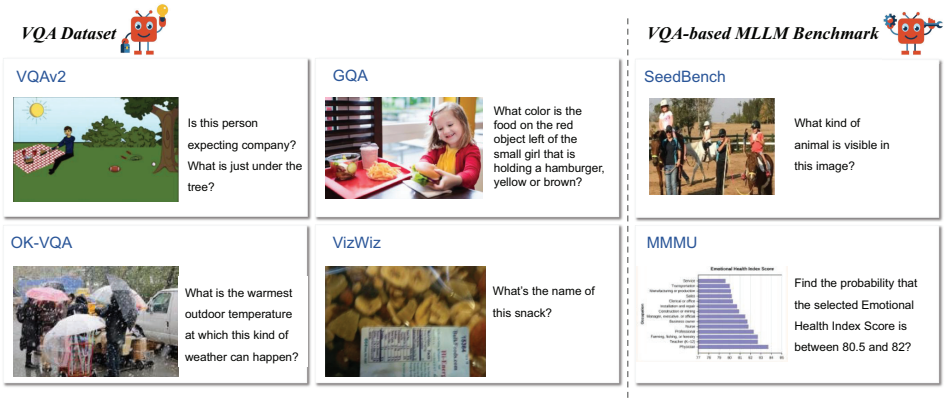


Fig. 9. Examples of widely-used VQA datasets and VqA-based MLLM benchmarks.

Table 6. Comparisons of Performance of Different Models on Four Widely-Used Datasets

Model name	Year	Dataset	VQA Accuracy	Model name	Year	Dataset	VQA Accuracy
mPLUG-Huge	2022	VQAv2 test-std	83.62	PaLI-X	2023	OK-VQA	66.1
Florence	2021	VQAv2 test-std	80.36	Prophet	2023	OK-VQA	62.5
LLaVa	2023	VQAv2 test-std	78.5(few-shot)	Flamingo	2022	OK-VQA	50.6
Flamingo	2022	VQAv2 test-std	67.6(few-shot)	PICa	2021	OK-VQA	48.0
BLIP-2	2023	VQAv2 test-std	65(zero-shot)	BLIP-2	2023	OK-VQA	45.9
IMG2LLM	2023	VQAv2 test-std	61.9(zero-shot)	MuKEA	2022	OK-VQA	42.59
UNITER	2019	VQAv2 test-std	73.4	KRISP	2021	OK-VQA	38.9
LXMERT	2019	VQAv2 test-std	72.5	ConceptBERT	2020	OK-VQA	33.66
VisualBERT	2019	VQAv2 test-std	71.0	LXMERT	2019	OK-VQA	32.04
Up-Down	2017	VQAv2 test-std	70.3	ViLBERT	2019	OK-VQA	31.35
DMN	2018	VQAv2 test-std	68.4	Mucko	2020	OK-VQA	29.2
MUTAN	2017	VQAv2 test-std	67.4	MUTAN	2017	OK-VQA	26.41
PaLI-X	2023	GQA test-dev	67.3	mOLUG-owl2	2023	VizWiz 2020	54.5(zero-shot)
LXMERT	2019	GQA test-dev	60.0	LLaVA	2023	VizWiz 2020	50.0(zero-shot)
BLIP-2	2023	GQA test-dev	44.7(zero-shot)	KOSMOS-1	2023	VizWiz 2020	35.3(few-shot)
TRRNet	2023	GQA Test2019	74.03	InstructBLIP	2023	VizWiz 2020	32.08(zero-shot)
VinVL	2023	GQA Test2019	64.85	Flamingo	2022	VizWiz 2020	49.8(few-shot)
BAN	2017	GQA Test2020	57.1	PaLI	2022	VizWiz 2020	73.3
Up-Down	2017	GQA Test2021	49.74	CLIP	2022	VizWiz 2020	61.64

settings across different models, particularly under limited or zero-shot scenarios, which can lead to substantial performance degradation relative to standard configurations. We hope these comprehensive comparisons provide researchers with deeper insights into model disparities and inspire future exploration in the field.

6 Problems and Challenges

We outline key problems and challenges in the current state of VQA that warrant further investigation.

Robustness to Data Bias. VQA models often learn from datasets with biases. Consequently, models frequently rely on statistical correlations rather than genuine understanding, leading to incorrect predictions when familiar patterns are absent. For instance, certain objects or scenes in a dataset may be disproportionately associated with specific answers, which undermines the model's ability to generalize to unseen or real-world data.

Explainability. VQA models often function as opaque black boxes, providing answers without clear justification. This lack of transparency raises concerns about their reliability, particularly in critical applications such as medical imaging or autonomous driving. Existing techniques, such as attention mechanisms, offer limited insight into the decision-making process. Improving explainability by providing human-interpretable reasoning is essential for building trust and facilitating adoption in sensitive domains.

Natural Language Generation. Most VQA systems rely on pre-defined answer libraries and prediction-based methods, which constrain their generative capabilities. While advancements in open-ended VQA research have been made, the ability to generate diverse, contextually accurate answers remains a significant challenge. Addressing this limitation is vital for enabling more natural and flexible interactions in VQA systems.

7 Conclusion

Since the VQA task was proposed, it has received great attention and gained rapid development and wide application. This survey presents a comprehensive review of the state-of-the-art on VQA task from two aspects. For the understanding of image-question pairs, feature extraction from individual modalities and information fusion methods between modalities are introduced, with particular attention to the application of visual-language pre-training models, and multimodal large language models. For the knowledge reasoning of graphical knowledge, we review knowledge sources and describe methods of acquisition and specific reasoning processes, highlighting differences in the type and difficulty of the models they include. Datasets of VQA and different evaluation metrics follow. We believe that the suggested possible future directions will benefit the specific task of VQA as well as the general objective of visual scene understanding.

References

[1] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. *CLEF (Working Notes)* 2, 6 (2019), 1–11.

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4971–4980.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* 35 (2022), 23716–23736.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[6] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*. Springer, 722–735.

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[8] Yang Bai, Ziran Li, Ning Ding, Ying Shen, and Hai-Tao Zheng. 2021. Infobox-to-text generation with tree-like planning based attention network. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 3773–3779.

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.

[10] Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. 2021. Visual question answering: Which investigated applications? *Pattern Recogn. Lett.* 151, C (Nov 2021), 325–331.
[11] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems* 26 (2013).
[12] Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Yaohang Li, Xing Luo, Chenyu Yi, and Alex C. Kot. 2024. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part L, volume 15108 of Lecture Notes in Computer Science*. 340–358.
[13] Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. 2021. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems* (2021).
[14] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023.
[15] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal LLM's referential dialogue magic. https://doi.org/10.48550/arXiv.2306.15195
[16] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al.. 2023. Pali-3 vision language models: Smaller, faster, stronger.
[17] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*. Springer, 104–120.
[18] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *CoRR*, abs/2404.16821. https://doi.org/10.48550/arXiv.2404.16821
[19] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.
[20] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.
[21] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.).
[22] Ana Claudia Akemi Matsuki de Faria, Felype de Castro Bastos, Jose Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. 2023. Visual question answering: A survey on techniques and common trends in recent literature. *CoRR*, abs/2305.11033, 2023.
[23] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 10088–10115.
[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
[25] Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. VQA: A new dataset for real-world VQA on PDF documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 585–601.
[26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5089–5098.
[27] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024. Association for Computational Linguistics*, 1107–1128.
[28] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 601–610.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net*, 2021. https://openreview.net/forum?id=YicbFdNTTy

[30] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18166–18176. https://doi.org/10.1109/CVPR52688.2022.01763

[31] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 8469–8488.

[32] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*. Springer, 363–370.

[33] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv:2306.13394. Retrieved from https://arxiv.org/abs/2306.13394

[34] Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. 2023. A challenger to gpt-4v? Early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436* (2023).

[35] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 457–468. DOI: https://doi.org/10.18653/v1/D16-1044

[36] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 6616–6628.

[37] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5067–5077.

[38] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? Dataset and methods for multilingual image question. *Advances in Neural Information Processing Systems* 28 (2015).

[39] Jingying Gao, Qi Wu, Alan Blair, and Maurice Pagnucco. 2024. Lora: A logical reasoning augmented dataset for visual question answering. *Advances in Neural Information Processing Systems* 36 (2024).

[40] Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven C. H. Hoi, and Xiaogang Wang. 2018. Question-guided hybrid convolution for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 469–485.

[41] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 489–498.

[42] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[43] Alex Graves. 2012. Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks* (2012), 37–45.

[44] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10867–10877.

[45] Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*. 14953–14962.

[46] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[48] 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. *arXiv preprint arXiv:2204.10448* (2022).

[49] Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. 2023. Let's think frame by frame with VIP: A video infilling and prediction dataset for evaluating video chain-of-thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 204–219.

[50] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10294–10303.

[51] Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7166–7176.

[52] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12976–12985.

[53] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020).

[54] Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6700–6709.

[55] Md. Farhan Ishmam, Md. Sakib Hossain Shovon, M. F. Mridha, and Nilanjan Dey. 2024. From image to language: A critical analysis of Visual Question Answering (VQA) approaches, challenges, and opportunities. *Information Fusion* 106 (2024), 102270. https://www.sciencedirect.com/science/article/pii/S1566253524000484

[56] Raisa Islam and Owana Marzia Moushi. 2024. GPT-4o: The cutting-edge advancement in multimodal LLM. *Authorea Preprints* (2024).

[57] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017* (2022).

[58] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10267–10276.

[59] Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. LLMs can find mathematical reasoning mistakes by pedagogical chain-of-thought. *CoRR* abs/2405.06705 (2024).

[60] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2901–2910.

[61] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4565–4574.

[62] Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding* 163 (2017), 3–20.

[63] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. 6769–6781.

[64] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[65] Jing Yu Koh, Daniel Fried, and Russ R. Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems* 36 (2024).

[66] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[67] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[68] Nguyen-Khang Le, Dieu-Hien Nguyen, Dinh-Truong Do, Chau Nguyen, and Le Minh Nguyen. 2024. Vietnamese elementary math reasoning using large language model with refined translation and dense-retrieved chain-of-thought. In *New Frontiers in Artificial Intelligence - JSAI International Symposium on Artificial Intelligence, JSAI-isAI 2024, Hamamatsu, Japan, May 28–29, 2024, Proceedings (Lecture Notes in Computer Science, Vol. 14741)*, Toyotaro Suzumura and Mayumi Bono (Eds.). Springer, 260–268.

[69] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. PMLR, 1188–1196.
[70] 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 *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3108–3120.
[71] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024. SEED-Bench-2-Plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790* (2024).
[72] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*.
[73] Dongxu Li, Junnan Li, and Steven Hoi. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems* 36 (2024).
[74] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.
[75] Guohao Li, Hang Su, and Wenwu Zhu. 2017. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. *arXiv preprint arXiv:1712.00733* (2017).
[76] Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1227–1235.
[77] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*. PMLR, 19730–19742.
[78] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086* (2022).
[79] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems* 34 (2021), 9694–9705.
[80] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10313–10322.
[81] Mingxiao Li and Marie-Francine Moens. 2022. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. *arXiv preprint arXiv:2203.02985* (2022).
[82] Qun Li, Fu Xiao, Bir Bhanu, Biyun Sheng, and Richang Hong. 2022. Inner knowledge-based Img2Doc scheme for visual question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 3 (2022), 1–21.
[83] Ruiyu Li and Jiaya Jia. 2016. Visual question answering with question representation update (qru). *Advances in Neural Information Processing Systems* 29 (2016).
[84] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409* (2020).
[85] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.
[86] Yanjie Li, Weijun Li, Lina Yu, Min Wu, Jingyi Liu, Wenqiang Li, Shu Wei, and Yusong Deng. 2024. MLLM-SR: Conversational symbolic regression base multi-modal large language models. *arXiv preprint arXiv:2406.05410* (2024).
[87] Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2023. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. *arXiv preprint arXiv:2311.11268* (2023).
[88] Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and S. Yu Philip. 2024. When LLMs meet cunning texts: A fallacy understanding benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
[89] Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, and Xunliang Cai. 2024. EAGLE: Elevating geometric reasoning through LLM-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397* (2024).
[90] Weixin Liang, Feiyang Niu, Aishwarya Reganti, Govind Thattai, and Gokhan Tur. 2020. LRTA: A transparent neural-symbolic reasoning framework with modular supervision for visual question answering. *arXiv preprint arXiv:2011.10731* (2020).
[91] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.

[92] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: You see what you think? or you think what you see? An image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566* (2023).

[93] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems* 36 (2024).

[94] Hugo Liu and Push Singh. 2004. ConceptNet–a practical commonsense reasoning tool-kit. *BT Technology Journal* 22, 4 (2004), 211–226.

[95] Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep mlp. *Patterns* 3, 7 (2022), 1–25.

[96] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[97] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32 (2019).

[98] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10437–10446.

[99] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems* 29 (2016).

[100] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. In *ICLR*.

[101] Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018. R-VQA: Learning visual relation facts with semantic attention for visual question answering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1880–1889.

[102] 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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28–December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.).

[103] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[104] Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS Datasets and Benchmarks*.

[105] Panzhong Lu, Xin Zhang, Meishan Zhang, and Min Zhang. 2022. Extending phrase grounding with pronouns in visual dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7614–7625. DOI: https://doi.org/10.18653/v1/2022.emnlp-main.518

[106] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8086–8098.

[107] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. 2018. Visual question answering with memory-augmented networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6975–6984.

[108] Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. 2024. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–20. DOI: https://doi.org/10.1109/TPAMI.2024.3366154

[109] Lin Ma, Zhengdong Lu, and Hang Li. 2016. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[110] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems* 27 (2014).

[111] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3195–3204.

[112] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1697–1706.
[113] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2200–2209.
[114] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 (2013).
[115] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 299–307.
[116] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in Neural Information Processing Systems* 31 (2018).
[117] Medhini Narasimhan and Alexander G. Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 451–468.
[118] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6087–6096.
[119] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 30–38.
[120] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. 2018. Learning conditioned graph structures for interpretable visual question answering. *Advances in Neural Information Processing Systems* 31 (2018).
[121] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.
[122] Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. TALM: Tool augmented language models. *CoRR* abs/2205.12255 (2022).
[123] Devshree Patel, Ratnam Parikh, and Yesha Shastri. 2021. Recent advances in video question answering: A review of datasets and methods. In *International Conference on Pattern Recognition*. Springer, 339–356.
[124] Cleon Pereira Júnior, Luiz Rodrigues, Newarney Costa, Valmir Macario Filho, and Rafael Mello. 2024. Can VLM understand children's handwriting? An analysis on handwritten mathematical equation recognition. In *International Conference on Artificial Intelligence in Education*. Springer, 321–328.
[125] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. 2021. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1753–1757.
[126] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
[127] Kiran Ramnath and Mark Hasegawa-Johnson. 2020. Seeing is knowing! fact-based visual question answering using knowledge graph embeddings. *arXiv preprint arXiv:2012.15484* (2020).
[128] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in Neural Information Processing Systems* 28 (2015).
[129] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst* 1, 2 (2015), 5.
[130] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28 (2015).
[131] Jonathan Roberts, Timo Lüddecke, Rehan Sheikh, Kai Han, and Samuel Albanie. 2023. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. *arXiv preprint arXiv:2311.14656* (2023).
[132] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. 2023. Visual chain of thought: Bridging logical gaps with multimodal infillings. *CoRR* abs/2305.02317 (2023).
[133] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.
[134] Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2017. High-order attention models for visual question answering. *Advances in Neural Information Processing Systems* 30 (2017).
[135] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6649–6658.
[136] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8876–8884.

[137] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14974–14983.

[138] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems* 36 (2024).

[139] Ying Shen, Min Yang, Yaliang Li, Dong Wang, Haitao Zheng, and Daoyuan Chen. 2023. Knowledge-based reasoning network for relation detection. *IEEE Trans. Neural Networks Learn. Syst.*, 34, 8 (2023), 5051–5063. (2021).

[140] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8376–8384.

[141] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4613–4621.

[142] Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2023. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *arXiv preprint arXiv:2310.00647* (2023).

[143] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*. Springer, 746–760.

[144] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[145] Harsimran Jit Singh, Gourav Bathla, Munish Mehta, Gunjan Chhabra, and Pardeep Singh. 2023. Visual questions answering developments, applications, datasets and opportunities: A state-of-the-art survey. In *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS'23)*. 778–785.

[146] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. 2021. Visual question answering using deep learning: A survey and performance analysis. In *International Conference on Computer Vision and Image Processing*. Springer, 75–86.

[147] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 539–559.

[148] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).

[149] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. 2018. Learning visual knowledge memory networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7736–7745.

[150] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[151] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[152] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[153] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[154] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C. H. Hoi. 2022. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7–11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 951–967.

[155] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.

[156] Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.

[157] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008.

[158] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 1290–1296.

[159] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 10 (2017), 2413–2427. https://doi.org/10.1109/TPAMI.2017.2754246

[160] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems* 36 (2024).

[161] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

[162] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. 2018. Chain of reasoning for visual question answering. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. 273–283.

[163] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *CoRR* abs/2303.04671 (2023).

[164] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163 (2017), 21–40.

[165] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4622–4630. https://doi.org/10.1109/CVPR.2016.500

[166] Xiaoqian Wu, Yong-Lu Li, Jianhua Sun, and Cewu Lu. 2023. Symbol-LLM: Leverage language models for symbolic system in visual human activity reasoning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems NeurIPS 2023*. http://papers.nips.cc/paper_files/paper/2023/hash/5edb57c05c81d04beb716ef1d542fe9e-Abstract-Conference.html

[167] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973* (2024).

[168] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. 2021. Probing intermodality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems* 34 (2021), 4514–4528.

[169] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441* (2023).

[170] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems* 36 (2024).

[171] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. 2021. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2197–2207.

[172] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3081–3089.

[173] 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 *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22–March 1, 2022*. AAAI Press, 3081–3089.

[174] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.

[175] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *CoRR* abs/2303.11381 (2023).

[176] Zhenguo Yang, Jiale Xiang, Jiuxiang You, Qing Li, and Wenyin Liu. 2023. Event-oriented visual question answering: The E-VQA dataset and benchmark. *IEEE Transactions on Knowledge and Data Engineering* 35, 10 (2023), 10210–10223.

[177] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR* abs/2311.04257 (2023). arXiv:2311.04257

[178] Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A. Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985* (2023).

[179] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3208–3216.

[180] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278* (2015).
[181] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 1821–1830.
[182] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al.. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of CVPR*.
[183] Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. 2025. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *IEEE Trans. Vis. Comput. Graph.* 31, 1 (2025), 525–535. https://doi.org/10.1109/TVCG.2024.3456159
[184] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023*. IEEE, 11941–11952.
[185] Dongxiang Zhang, Rui Cao, and Sai Wu. 2019. Information fusion in visual question answering: A survey. *Information Fusion* 52 (2019), 268–280. https://www.sciencedirect.com/science/article/pii/S1566253518308893
[186] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. 2024. CMMMU: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944* (2024).
[187] Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024. Benchmarking large multimodal models against common corruptions. In *NAACL*.
[188] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624* (2024).
[189] Xin Zhang, Wen Xie, Ziqi Dai, Jun Rao, Haokun Wen, Xuan Luo, Meishan Zhang, and Min Zhang. 2023. Finetuning language models for multimodal question answering. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM'23)*. Association for Computing Machinery, New York, NY, USA, 9420–9424. https://doi.org/10.1145/3581783.3612837
[190] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *CoRR* abs/2302.00923 (2023).
[191] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: Empowering vision-language model with multi-modal in-context learning. *CoRR* abs/2309.07915 (2023). arXiv:2309.07915
[192] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).
[193] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4995–5004.
[194] Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2021. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Yokohama, Yokohama, Japan) *(IJCAI'20)*. Article 153, 7 pages.
[195] Maryam Ziaeefard and Freddy Lecue. 2020. Towards knowledge-augmented visual question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1863–1873.
[196] Yeyun Zou and Qiyu Xie. 2020. A survey on VQA: Datasets and approaches. In *2020 2nd International Conference on Information Technology and Computer Application (ITCA'20)*. IEEE, 289–297.

Received 13 January 2023; revised 19 October 2024; accepted 14 December 2024