Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges

BYEONG SU KIM, Yonsei University, Seoul, Korea (the Republic of)
JIEUN KIM, Yonsei University, Seoul, Korea (the Republic of)
DEOKWOO LEE, Keimyung University, Daegu, Korea (the Republic of)
BEAKCHEOL JANG, Yonsei University, Seoul, Korea (the Republic of)

Visual question answering (VQA) is a dynamic field of research that aims to generate textual answers from given visual and question information. It is a multimodal field that has garnered significant interest from the computer vision and natural language processing communities. Furthermore, recent advances in these fields have yielded numerous achievements in VQA research. In VQA research, achieving balanced learning that avoids bias toward either visual or question information is crucial. The primary challenge in VQA lies in eliminating noise, while utilizing valuable and accurate information from different modalities. Various research methodologies have been developed to address these issues. In this study, we classify these research methods into three categories: Joint Embedding, Attention Mechanism, and Model-agnostic methods. We analyze the advantages, disadvantages, and limitations of each approach. In addition, we trace the evolution of datasets in VQA research, categorizing them into three types: Real Image, Synthetic Image, and Unbiased datasets. This study also provides an overview of evaluation metrics based on future research directions. Finally, we discuss future research and application directions for VQA research. We anticipate that this survey will offer useful perspectives and essential information to researchers and practitioners seeking to address visual questions effectively.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Multimedia and multimodal retrieval**; **Question answering;**

Additional Key Words and Phrases: Visual question answering, multi-modal, attention mechanism, model-agnostic, language bias, computer vision, natural language processing

ACM Reference Format:
Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges. *ACM Comput. Surv.* 57, 10, Article 249 (May 2025), 35 pages. https://doi.org/10.1145/3728635

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government under Grant RS-2023-00273751.

Authors' Contact Information: Byeong Su Kim, Yonsei University, Seoul, Korea (the Republic of); e-mail: kbs23@yonsei.ac.kr; Jieun Kim, Yonsei University, Seoul, Korea (the Republic of); e-mail: lilly9928@yonsei.ac.kr; Deokwoo Lee, Keimyung University, Daegu, Korea (the Republic of); e-mail: dwoolee@kmu.ac.kr; Beakcheol Jang (Corresponding author), Yonsei University, Seoul, Korea (the Republic of); e-mail: bjang@yonsei.ac.kr.

ACM 0360-0300/2025/05-ART249
https://doi.org/10.1145/3728635

1 Introduction

Visual Question Answering (**VQA**) is a growing research area within the broader multimodal AI field, integrating **computer vision** (**CV**) and **natural language processing** (**NLP**) to answer textual questions about images. In some cases, VQA incorporates external knowledge alongside images and questions to enrich the understanding and generation of answers. Recent research has also explored combining vision, NLP, and audio, making VQA a crucial foundation for future AI development. Significant advancements in CV, including image classification [37], segmentation [112], and object detection [129], as well as NLP innovations like speech recognition [126] and machine translation [11], have contributed to the progress of VQA. This progress has led to applications in learning assistance systems, medical image analysis [92], and support for the visually impaired [51]. Multimodal research, which combines different data types such as images, videos, text, and audio, is tackling various societal challenges. VQA, as a complex and interdisciplinary task, involves extracting and fusing features from vision and language, often leveraging common sense and external knowledge to provide accurate answers. In this sense, VQA can be considered a fundamental challenge that the "AI-complete" task [8].

For VQA tasks, the objective is to predict the correct answer by fusing the information from a given image and a related question within a dataset. The challenge with VQA lies in eliminating noise, while utilizing valuable information when extracting features from different modalities [13]. Accordingly, several methods have been proposed for addressing this issue. In this survey, we divided the methods into three groups: **Joint Embedding** (**JE**), **Attention Mechanism** (**AM**), and **Model-agnostic** (**MA**) methods.

VQA datasets are categorized into three types: Real Image [8, 31, 51, 104, 108, 113, 142, 152], Synthetic Image [8, 69, 101, 109, 131, 141, 144], and Unbiased Datasets [3, 45, 65]. Real Image datasets, such as VQA [8] and VizWiz [51], capture diverse real-world scenarios, making them useful for generalized applications; however, their noisy nature often hampers accurate inference. Synthetic Image datasets, like CLEVR [69], are designed to reduce noise and focus on evaluating spatial and logical relationships, enabling precise assessment of inference capabilities. Unbiased Datasets, such as VQA-CP [3] and VQA-v2 [45], address dataset bias by using different answer distributions in training and test sets, which discourages models from relying solely on superficial correlations in question-answer pairs. Many existing models, when tested on these datasets, were found to depend heavily on dataset biases rather than true reasoning. Furthermore, evaluation metrics now extend beyond traditional VQA accuracy to include grammatical and semantic assessments, promoting the creation of models capable of generating richer and more comprehensive answers.

The remainder of this article is organized as follows: Section 2 provides a general background on the VQA task by dissecting the terms V, Q, and A to understand the VQA task. It also introduces various related surveys in the field of VQA. Section 3 presents an overview of the JE method, initially proposed as a VQA model. It discusses its limitations and describes other models that have evolved from this approach. Section 4 introduces the AM method, which focuses on learning features crucial for generating correct answers, without simply combining features. It also presents models that employ this approach. In Section 5, we explore the MA method to solve the limitations of dataset distribution and model training. Section 6 analyzes the various proposed datasets to validate the VQA models, and outlines the purpose, advantages, and disadvantages of the datasets. In Section 7, we discuss the problems and limitations of VQA based on the current research, while exploring future research directions and practical applications. Finally, Section 8 concludes the article, summarizing key findings and highlighting the significance of VQA in the broader context of multimodal AI research.

Q : Are there shadows on the sidewalk?
A : Yes

(a) Yes/No

Q : What kind of bears are these? A : 1
1. Brown 2. Grizzly 3. Polar

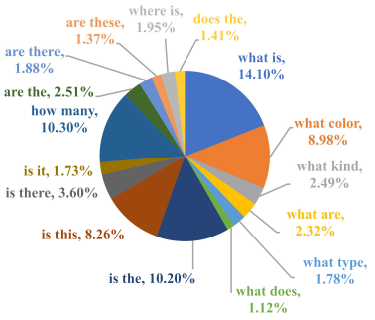
(b) Multi-choice

Q : How many windows can you see?
A : 1

(c) Counting

Q : What is the man doing in the street?
A : Walking

(d) Open-ended



(e) The first 2 words of the question's distribution.

Fig. 1. (a)–(d) Examples of representative answer types from the VQA dataset and (e) Share of the first two letters of the question in the VQA v1 [8] dataset.

2 Background Knowledge and Related Works

2.1 Background Knowledge

VQA is a task that receives visual information and questions as inputs and outputs the correct answers. To comprehend this task thoroughly, we delve into the characteristics of each component: Visual, Question, and Answer.

Visual information typically encompasses both images and videos. These mediums contain information on diverse scenarios essential for training models to answer a wide range of creative questions. Most images and videos contain information of everyday life scenes. However, in Ref. [8], there are also data that create an abstract situation by utilizing a "paperdoll". In the case of video, in addition to those portraying daily life, there are also videos on non-daily activities featuring cinematic special effects. Notably, this survey primarily focused on image QA, with more detailed datasets presented in Section 6.

Questions were posed in reference to a given image. For example, in Figure 1(d), the question, "How many people are in the image?" can be obtained by inferring the answer from the image alone. However, for questions like, "What brand is the food truck in the image and what year is it?" requires external knowledge. Some datasets include questions demanding external knowledge, adding an extra layer of complexity. The question format comprises different types of questions to test the generality and robustness of the model. Figure 1(e) shows examples that constitute over 1% of the first two letters of the question in the VQA v1 dataset [8]. There are a wide variety of questions.

Answer were divided into two types: classified and open-ended categories. Classification typically involves counting, multiple-choice, and yes or no problems. In such cases, we identify the class with the highest probability of being the correct answer. To realize this, softmax is used for deep learning. For open-ended problems, a decoder-based **recurrent neural network** (**RNN**) was used to generate sentences. Examples are shown in Figure 1(a)–1(d).

Table 1. Related Survey

Study	Year	Scope
Visual Question Answering: A Survey of Methods and Datasets [154]	2017	Early methods for VQA, such as JE and AMs, are reviewed. Neural Module Networks (NMN) and Dynamic Memory Networks (DMN) are then introduced to address the limitations associated with the monolithic nature of CNN and RNN.
Information fusion in visual question answering: A Survey [166]	2019	Information fusion between visual and textual channels plays an important role in improving overall accuracy. There are various methods proposed for fusion, and an abstract fusion framework that can be applied to major VQA models is introduced.
A survey of methods, datasets and evaluation metrics for visual question answering [134]	2021	Models of different AMs were divided into single-hop and multi-hop. Traditional VQA models as well as models with the ability to read text provided in images were investigated.
Language bias in Visual Question Answering: A Survey and Taxonomy [165]	2021	To address language bias, existing methods are organized into three categories. The relevant representative methods in each category are introduced, summarized, and analyzed, and the causes of language bias are introduced and classified. As a dataset, the VQA-CP dataset [3] is introduced, which allows for limited experimentation on language bias.
The multi-modal fusion in visual question answering: a review of attention mechanisms [102]	2023	Provides an overview and description of an AM based model that performs well on VQA tasks. It also provides visualizations of various analyses using CiteSpace [128].

Text-Based Question Answering has been a significant inspiration for VQA. The Q&A system is an area of active research in NLP that dates back to the 1960s [46] and continues to this day [165]. The Q&A system aims to answer any given question in a specific context. Compared with search engines, it maximizes user convenience by providing the final answer to the user's question rather than returning a list of hyperlinks in response to the query. Major technology companies like Google have also attempted to provide high-quality user experiences by introducing Q&A techniques [38]. Recently, the "Retriever-Reader" approach, which comprises a "Retriever" to search for relevant documents in the Database and a "Reader" to answer questions from the retrieved documents, has been studied [52].

2.2 Related Works

Since [8], multiple models and datasets have been proposed for VQA tasks, with researchers having analyzed and evaluated VQA models and datasets from their own perspectives. Table 1 provides a summary of the relevant surveys.

Wu et al. [154] examined state-of-the-art methods using a recent model. They categorized all the approaches into JE methods, AMs, and compositional models based on how they connect visual and textual features. They also discussed memory-augmented and modular architectures that use structured knowledge bases. The datasets employed included natural images with

general situation data, clipart images with non-naturally synthesized images, and knowledge-based enhanced datasets that require external knowledge.

Zhang et al. [166] stated that the VQA task remains challenging because it requires an understanding of the semantic information in the textual and visual channels. A general solver for VQA tasks involves feature extraction, feature fusion, and answer prediction. The fusion of visual and question information was emphasized as a crucial aspect of improving accuracy. Fusion methods can be categorized as vector operators, deep neural networks, bilinear pooling, AMs, and memory networks. In this survey, the composition of the fusion techniques of the VQA domain proposed thus far was organized, and the mainstream VQA model was organized using the abstract fusion framework proposed in the survey.

Sharma et al. [134] summarized some of the fundamental concepts of VQA systems and the previous efforts made to address the core concept problem. In addition to traditional VQA models, they investigated models with the ability to read text provided in images, and discussed the TextVQA [79], **Scene Text Visual Question Answering** (**ST-VQA**) [16], and OCR-VQA [111] datasets developed to evaluate these models, as well as the newly developed datasets GQA [65] and **outside knowledge visual question answering** (**OK-VQA**) [108] in 2019 and 2020.

Yuan Desen [165] argues that VQA suffers from a language bias problem, which reduces the robustness of the model and has a negative impact when used in realistic situations with model. Notably, the author conducted the first survey in this field. The existing methods were divided into three categories: enhancing visual information, weakening language priors, data enhancement, and training strategies. Relevant representative methods in each category were introduced, summarized, and analyzed. Thus, the causes of the language bias were identified and classified. In addition to the commonly used VQA dataset, the VQA-CP dataset [3], which allows for limited experiments on language bias, was described, with the experimental results of various methods summarized.

Lu et al. [102] argued that the key task in achieving multimodal fusion in VQA is the AM. Therefore, this survey describes the need for the development of AMs, provides an overview of current AM methods that exhibit good performance, analyzes their shortcomings, describes ways to improve them, and suggests future research directions. The authors also suggested that the introduction of factor-based a priori conditions or knowledge graphs can significantly improve the accuracy and interpretability of VQA tasks. This survey provided a detailed analysis of various VQA models based on AMs.

Notably, there have been several surveys on VQA. The survey in Refs. [134, 154] mainly focused on classifying and overviewing VQA models, but also categorized and introduced datasets. The surveys in Refs. [102, 166] focused on AMs and provided detailed analysis. Finally, the survey in Ref. [165] emphasized on language bias, briefly introducing and classifying the models that attempted to solve it. By contrast, this survey classified the models into three categories: JE, AMs, and recent research as MAs. We also analyzed the advantages and disadvantages of each research direction and made suggestions for VQA research based on current findings.

3 Joint Embedding

JE of images and text has been attempted in the field of image captioning [36]. The goal of image captioning is to automatically generate descriptions of images in a manner that humans can understand [76]. This requires CV-related knowledge regarding the extraction of features from an image. Additionally, NLP-related expertise is also required for generating natural language–a language that humans can understand (from the features of an image). Image captioning has benefited from the rapid growth in CV and NLP fields. Figure 2 depicts the abstract framework of the JE method, and a summary of the models used in this survey is presented in Table 2.

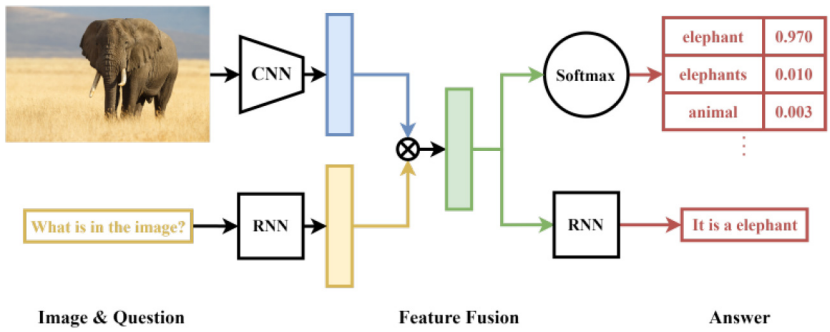


Fig. 2. An abstract framework for JE methods.

Table 2. Overview of VQA Models

	Model	Joint Embedding	Attention Mechanism	Model Agnostic	Year	Question Features	Image Features
Joint Embedding	VanillaVQA [8]	✓			2015	LSTM	VGGNet
	Neural-Image-QA [105]	✓			2015	LSTM	GoogLeNet
	mQA [41]	✓			2015	LSTM	GoogLeNet
	MCB [40]	✓			2016	LSTM	VGGNet, ResNet
	MLB [78]	✓			2016	GRU	ResNet
	BAN [77]	✓	✓		2018	GloVe, GRU	Faster R-CNN
	MDFNet [169]	✓	✓		2021	Bi-GRU	Faster R-CNN, ResNet
	N-KBSN [103]	✓	✓		2021	ELMo	Faster R-CNN
	DMBA-NET [158]	✓	✓		2022	BERT	Faster R-CNN
	VB-MVQA [19]	✓	✓		2023	GloVe, LSTM	ResNet
Attention Mechanism	Word+Region Sel. [137]	✓	✓		2016	word2vec	VGGNet
	SAN [160]	✓	✓		2016	LSTM, CNN	VGGNet
	Up-Down [7]	✓	✓		2018	LSTM	Faster R-CNN
	NS-CL [106]	✓	✓		2019	-	Mask R-CNN, ResNet
	SelRes [58]	✓	✓		2020	GloVe	Faster R-CNN
	UFSCAN [168]	✓	✓		2021	GRU	Faster R-CNN
	LSAT [136]	✓	✓		2022	GloVe, GRU	Faster R-CNN, ResNet
	VLMo [14]	✓	✓		2022	BERT	ViT
	CSCA [110]	✓	✓		2023	GloVe	Faster R-CNN, ResNet
	ACOCAD [9]	✓	✓		2024	GloVe, LSTM	Faster R-CNN
Model Agnostic	RUBi (MuRel) [18]			✓	2019	GRU	Faster R-CNN
	Learned-Mixin (Up-down) [34]			✓	2019	LSTM	Faster R-CNN
	Learned-Mixin+H (Up-down) [34]			✓	2019	LSTM	Faster R-CNN
	VGQE (MuRel) [81]			✓	2020	GloVe	Faster R-CNN
	CSS (LM+H) [25]			✓	2020	LSTM	Faster R-CNN
	CIKD (Up-Down) [118]			✓	2021	LSTM	Faster R-CNN
	WMA (MLB and MCB) [86]			✓	2021	-	WMA R-CNN
	GGE (Up-Down) [53]			✓	2021	GloVe	Faster R-CNN
	PromptCap (OFA) [61]			✓	2022	-	-
	BEiT-3 (VLMo) [150]			✓	2023	-	-
	Prismer (Mask2Former and 5 others) [97]			✓	2023	RoBERTa	ViT
	CuMo (CLIP and Mistral) [85]			✓	2024	-	CLIP ViT
	Florence-VL (Florence-2 and Llama-3) [24]			✓	2024	-	Florence-2

3.1 Vanilla VQA

Antol et al. [8] proposed the Vanilla VQA model, which employs the last hidden layer of the VGGNet model of CNN to encode the image data. The model was then constructed using an RNN model, LSTM [57]. This was to extract a representation of the question and combine it with the encoded image feature map to predict the correct answer. Antol et al. proposed a VQA dataset based on the **Microsoft Common Objects in COntext** (**MS-COCO**) dataset [90] along with this model and conducted experiments using the proposed dataset. When tested on the proposed dataset, it demonstrated 57.75% accuracy for narrative questions and 62.70% accuracy for multichoice questions.

3.2 Neural Image QA

Malinowski et al. [105] proposed a method that uses a CNN and an LSTM. They used a CNN pre-trained with the ImageNet dataset to extract image features. The LSTM unit contains an embedding of each word along with the extracted image features. Encoding stops and decoding starts when the last token of the question "?" is encountered. At this point, the embedded features of the question and image are used to iteratively generate a sentence until the END symbol is predicted. This shows that RNNs can be used to solve the open-ended problem of predicting answers of variable lengths. In the experimental section, the model is as follows: 19.43% accuracy on the single word dataset from **DAtaset for QUestion Answering on Real-world images** (**DAQUAR**) [104].

3.3 Multimodal QA

Gao et al. [41] proposed **Multimodal QA** (**mQA**), which combines a CNN and an RNN. mQA uses LSTM [57], an RNN model, to extract the semantic representation of a question, and a CNN to extract the image representation. Subsequently, LSTM [57] extracts the current word expression and linguistic context from the answers, with the model constructed by integrating the three extracted expressions to generate the next word for the final answer. This model is similar to that proposed by Malinowski et al. [105]. However, because the weights of the LSTM are not shared by the encoder and decoder, the question can be better understood, the quality of the answer can be improved, and image features can be used to generate the correct answer instead of encoding image features, such as the question. When evaluated on the FM-IQA dataset, a normal human demonstrated 94.8% accuracy for correct answers, while mQA demonstrated 64.7% accuracy.

3.4 Multimodal Low-Rank Bilinear

Kim et al. [78] proposed a **Multimodal Low-rank Bilinear** (**MLB**) attention network based on the Hadamard products. Bilinear pooling generates high-dimensional features that have limitations due to their high computational cost. Therefore, in a previous study [43], an attempt was made to reduce the expansion of dimensionality while maintaining performance. However, MLB has a more flexible structure because it uses linear mapping and the Hadamard product, replacing MCB. When comparing the performance with MCB, the VQA test-standard benchmark showed accuracies of 66.47% and 66.89% with MCB and MLB, respectively.

3.5 BAN

Kim et al. [77] also proposed a **bilinear attention network** (**BAN**). Previously, co-attention networks [100, 114] performed well in inferring visual and textual attention distributions for each modality. However, co-attention networks only selectively attend to question words along with a portion of the image regions and ignore the interaction between words and visual regions to avoid increasing computational complexity. The BAN proposed by the authors extends the co-attention concept to bilinear attention, which considers all pairs of question words and image regions. It is based on MLB [78], with the Hadamard product as the effective AM.

3.6 MDFNet

Zhang et al. [169] argued that spatial and contextual relationships are important for image QA models to predict the correct answers. In this study, inspired by the success of image representation in graph form, as in Ref. [26], the authors designed a **Graph Reasoning and Fusion Layer** (**GRFL**) and proposed a **Multimodal Deep Fusion Network** (**MDFNet**) stacking GRFL. GRFL is composed of a question encoder, question-guided visual attention, dual graph reasoning, and bilinear fusion. In dual-graph reasoning, the authors use semantic and spatial graphs, which are

the foundational premise of their approach, to infer the semantic and spatial relationships between questions and images. In the experimental section, MDFNet exhibited 71.19% accuracy on the VQA v2 test-dev, 71.32% accuracy on the test standard, and 57.05% accuracy on GQA.

3.7 N-KBSN

Ma et al. [103] proposed a JE model based on a dynamic word vector to solve the limitation of existing JE-based models using static word vectors for text characterization, which do not reflect the real-world language environment. This is referred to as a **non-KB-specific network** (**N-KBSN**) model, which is divided into three main parts: question text and image feature extraction; self-attention and guided attention; and feature fusion and classification. An important feature of this model is the use of ELMo and feature enhancement based on the multihead AM for text characterization. With this structure, the proposed model can effectively handle polysemy, which is difficult to solve using static vectors. In the experimental section, the N-KBSN(l) model with $ELMo_l$ and a word vector dimension of 1,024 showed 67.72% accuracy for the VQA v2 val.

3.8 DMBA-NET

Yan et al. [158] designed two basic attention units, BAN-GA and BAN-SA, inspired by MCAN [164], which simultaneously explored inter-modality and intra-modality relations. They are based on the dynamic word vector of BERT to encode questions and self-attention to further process the question features. Unlike other models, they used bilinear attention instead of the dot product to calculate the inter-modality and intra-modality attention to reduce the input dimension and computation. In the experimental section, DMBA-NET showed an accuracy of 69.45% for the VQA v2 val.

3.9 VB-MVQA

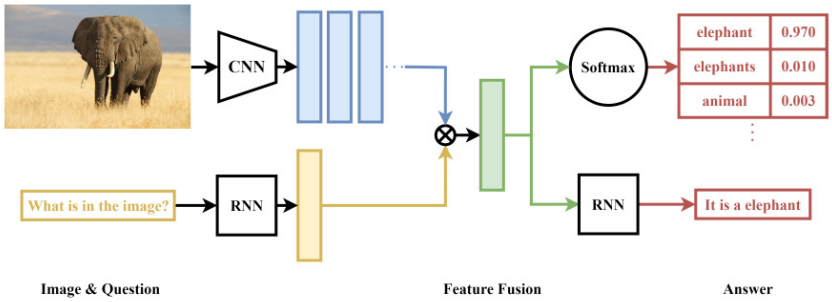
Cai et al. [19] proposed the use of a bilinear approximation for VQA in the medical domain. In general, VQA tasks suffer from language bias, which worsens when the data are sparse. Therefore, the authors introduce a pre-training method for medical image feature extraction based on **contrastive language image Pre-training** (**CLIP**) [125] and AMs to alleviate the problem. They propose **vision-conditioned reasoning and bilinear attention for the MedVQA** (**VB-MVQA**) model to further enhance the semantic information of images in the testing phase. VB-MBQA was 71.3% accurate for **visual question answering in radiology** (**VQA-RAD**), and 78.7% accurate for semantically-labeled knowledge-enhanced dataset for medical visual question answering (SLAKE).

4 Attention Mechanism

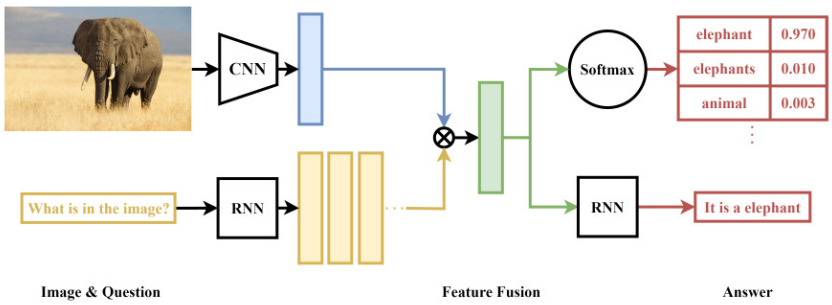
For JE, both image and question information are embedded into one feature; therefore, the information that is not related to the correct answer is also embedded. However, because it is usually important to focus on certain parts of the image related to the answer or specific words in the question to answer the question, irrelevant information becomes noisy and degrades the performance. To address the limitations of JE, methods have been proposed that use AMs to focus on the local features of the image required to generate the correct answer and focus on important words in the question. Figure 3 depicts the abstract framework of the AM. Next, we describe a representative VQA model using an AM.

4.1 Word+Region Sel

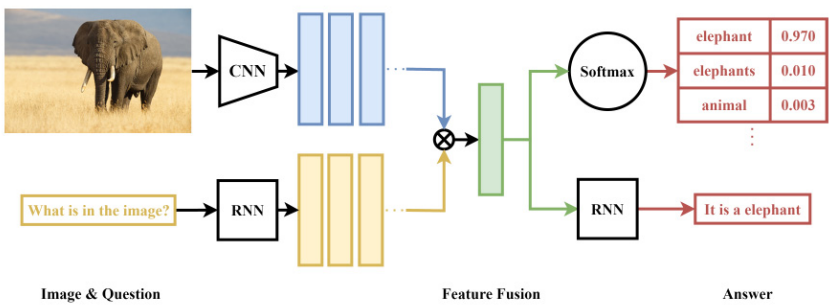
Shih et al. [137] addressed the challenge of answering natural language questions about images by proposing a model that selects relevant image regions. They used the top 99 EdgeBoxes with non-maximum suppression (0.2 IoU overlap) to identify small regions related to the question, adding



(a) An abstract framework for the visual attention model.



(b) An abstract framework for the question attention model.



(c) An abstract framework for the co-attention model.

Fig. 3. An abstract framework for AM methods.

the entire image as an additional region for a total of 100 candidate regions per image. Each region was represented with a 5096-dimensional vector, combining the VGG-s network's [23] last fully connected layer (4096 dimensions) and pre-softmax layer (1000 dimensions) outputs. Their method outperformed an LSTM-based model on the VQA v1 test standard, achieving 62.44% accuracy compared with 57.17%.

4.2 SAN

Yang et al. [160] queried the semantic representation of a question to find image regions associated with the correct answer and proposed **stacked attention networks (SANs)** that can incrementally infer the correct answer by querying the image multiple times. SAN can be viewed as an

extension of the AM, which has been successfully applied in image captioning and machine translation [11]. In this study, we showed that repeatedly using an AM for VQA can improve the ability to focus on image regions that are relevant to the correct answer. They performed experiments with one or two attention layers on the DAQUAR, COCO-QA, and VQA datasets. In DAQUAR-ALL, an SAN with two layers showed 29.3% accuracy; in COCO-QA, it showed 61.6% accuracy; and in the VQA dataset, it showed 58.7% accuracy.

4.3 Up-Down

Anderson et al. [7] introduced the bottom-up AM for VQA, addressing the limitation of top-down attention, which focuses only on task-specific areas. Bottom-up attention, inspired by the human visual system, automatically focuses on salient regions in the image. Unlike the SAN [160] model, their up-down model computes attention at both the object and salient region levels using Faster R-CNN [129] for object detection. Their approach achieved record scores in image description evaluation and VQA tasks, setting a new benchmark with 70.34% accuracy on the VQA challenge in 2017. This model has since become a foundational approach in VQA research.

4.4 NS-CL

Mao et al. [106] explored interpretable object-based visual representations, such as **neural-symbolic visual question answering** (**NS-VQA**) [161]. Unlike NS-VQA, which performs well, but requires fully annotated scenes during training, **Neuro-Symbolic Concept Learner** (**NS-CL**) learns its representation based solely on natural supervision (questions and answers). In addition, in Up-Down [7], the authors proposed representing images as a collection of convolutional object features that significantly improved VQA. The Up-Down [7] model encodes questions using a neural network and answers them through question-conditioned attention to object features. By contrast, NSCL parses question inputs into programs and executes them on object features to obtain answers. This rendered the reasoning process more interpretable. Thus, NS-CL enables interpretable, robust, and accurate visual inference. It exhibits an accuracy of 98.9% for the CLEVR dataset. NS-CL is novel in that it solves VQA problems based on neurosymbolic AI ideas.

4.5 SelRes

Hong et al. [58] proposed **Selective Residual Learning** (**SelRes**) to enhance VQA models by focusing on important relations relevant to the correct answer. SelRes employs residual learning [55] to distinguish significant relationships and includes a Selective Residual module that highlights image regions related to the answer. Additionally, they introduced selective masking, which filters attention maps based on vector importance. Fine-tuning LXMERT [143] with SelRes on the VQA v2 dataset achieved 72.7% and 72.8% accuracy on the test-std set. Although SelRes applied to question self-attention underperformed, resolving this issue could improve outcomes.

4.6 UFSCAN

Zhang et al. [168] proposed the UFSCAN model for VQA, addressing the limitation of previous models [100, 114, 137, 160] that focused only on spatial attention (where to look). They emphasized the importance of semantic feature attention (what to look at) and introduced a **multimodal feature-wise attention module** (**MulFA**) for both images and questions. UFSCAN combines this with a feature-wise co-attention module (MulFCoA), a **visual-spatial attention module** (**VSA**), and a **multimodal residual module** (**MulRM**) to fuse visual and textual information effectively. Their experiments on VQA v1 and VQA v2 achieved significant performance, with UFSCAN scoring over 70% on both datasets.

4.7 LSAT

Shen et al. [136] found that it is difficult to achieve a good performance in VQA tasks using a traditional transformer model, which only models global self-attention, and proposed a novel local self-attention scheme. This scheme is called **Local Self-Attention in Transformer** (**LSAT**). The LSAT model can capture rich contextual visual information features by modeling intra- and inter-window attention simultaneously by setting local windows for visual features. They used visual grid features to solve the global problem in transformers. In the experimental section, the authors experimented on VQA v2 and CLEVR. They demonstrated performances of 98.72% on CLEVR, 71.94% on test-std, and 71.67% on Test-dev of VQA v2.

4.8 VLMo

Bao et al. [14] introduced VLMo, a Vision-Language pre-trained model designed for universal use across vision-language tasks. Unlike previous models, which are either dual-encoder-based (e.g., CLIP [125], ALIGN [68]) or fusion-encoder-based, VLMo combines the strengths of both architectures. It uses a **Mixture-of-Modality-Experts** (**MoME**) transformer, with separate experts for vision, language, and vision-language tasks, allowing for efficient alignment of vision and language information. VLMo was pre-trained on large datasets like Conceptual Captions [135], MS-COCO [90], and Visual Genome [80]. It achieved high performance, scoring 82.88% on VQA v2 test-dev and 88.62% on NLVR2 dev.

4.9 CSCA

Mishra et al. [110] proposed the **Cascade of Self- and Co-Attention** (**CSCA**) model, combining **self-attention** (**SA**) for contextual information and **co-attention** (**CA**) for interaction between images and text. The model integrates these advantages into a single self- and co-attention-based attention block (SCA), which processes both text and image modalities. By stacking multiple SCA blocks, CSCA extracts fine-grained information. Experiments on the VQA v2 and TDIUC [70] datasets showed 88.12% performance on TDIUC and 67.36% on VQA v2. Increasing the number of SCA blocks beyond four did not improve performance further.

4.10 ACOCAD

Asri et al. [9] point out that models using traditional AMs suffer from the problem of excessive interaction, which can introduce noisy information into the model. Therefore, the authors propose an Advanced Visual and Textual Co-context Aware Attention Network with Dependent Multimodal Fusion Block for VQA (ACOCAD), which integrates question-level and word-level visual AMs to eliminate unnecessary interactions between words in a question and interactions between words and image regions, and adds the **Universal Sentence Encoder** (**USE**) [20] model to improve the word AM in the process of extracting text features. Finally, since **independent multimodal fusion** (**IMF**), which is commonly used to answer the question, ignores the dependence between regions and words, the authors propose a **dependent multimodal fusion** (**DMF**) block, which considers words and regions simultaneously and modifies the weights and word vectors according to image regions. In the experiments, ACOCAD performed 71.18% on VQAv2 test-std and 57.37% on GQA test-dev. However, it used the sigmoid function to select the final answer, which has limitations in learning complex decision boundaries and needs to be improved.

5 Model Agnostic

In recent years, VQA research has focused on fixing aspects other than the model by expressing them in an MA framework or training scheme, rather than modifying the model. Therefore, models

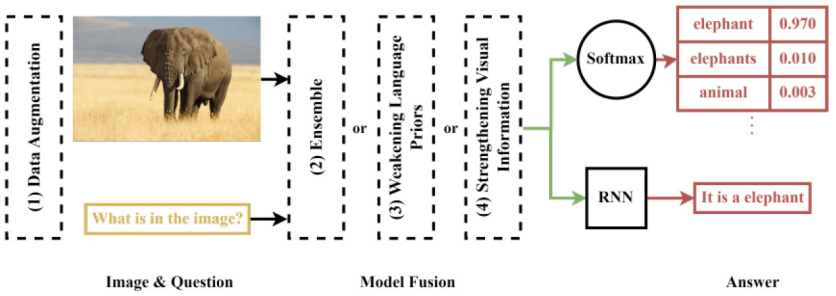


Fig. 4. An abstract framework for MA methods.

using JE and AM methods demonstrated good performance in VQA tasks. However, problems such as dataset bias, shortcut bias, visual explainability, and question sensitivity, which are currently the most significant problems in VQA, are generated during model learning or data distribution [165]. To this end, researchers have been working on methods that do not modify the model, but rather (1) learn by augmenting data [25], (2) use ensembles [97, 118, 150], (3) use the strategy of weakening language priors [18, 34, 53], and (4) use the strategy of strengthening visual information [61, 81, 86]. We refer to these studies as MA methods and the base model of the proposed method is provided in parentheses. Figure 4 illustrates the abstract framework of the MA method. Next, we describe the results of our recent study.

5.1 RUBi (MuRel)

Cadene et al. [18] proposed RUBi, a strategy for reducing unimodal biases in VQA. They found that VQA models often exploit statistical regularities [2, 4, 65, 69], like associating "yellow" with "banana," instead of analyzing the image. RUBi addresses this by dynamically adjusting the loss function to reduce bias, strengthening learning on less biased examples. In their experiments on VQA-CP v2 and VQA v2, RUBi improved performance across various VQA architectures, achieving 47.11% on VQA-CP v2 and 63.18% on VQA v2 validation. While RUBi is effective, it relies on a question-only model to mitigate bias, which is a limitation.

5.2 Learned-Mixin (Up-Down)

Clark et al. [34] referred to bias as a pattern of selecting a certain percentage of words in a VQA task or text close to a word in a question as an answer. The authors intentionally learn about this bias and ensemble the biased model with a second model to prevent the second model from using the bias so that it performs well, even on unbiased datasets. For this purpose, they proposed the **LearnedMixin** (**LM**) and **Learned-Mixin +H** (**LMH**) methods. For the LM, the model can determine the reliability of a given bias. The LMH adds an entropy penalty to the loss to prevent it from ignoring the bias. In their experiments, the authors experimented on VQA-CP v2 and found that the LMH model with up-down as the base performed at 52.05%.

5.3 VGQE (MuRel)

Kv et al. [81] proposed a new MA question encoder, the **visually grounded question encoder** (**VGQE**), to strengthen visual information. This prevents the VQA models from being biased by language priors during training. The VGQE uses the RNN-based question-encoding approach as a base and modifies it to find visual features that are relevant to the current question word to generate visually grounded word embeddings. To experiment with it, the authors used a simplified MUREL VQA model as the base model, and the results showed 50.11% accuracy on VQA-CP v2. While

models that performed well on the VQA-CP dataset showed performance degradation on the VQA dataset, VGQE showed an unusual result, with the baseline showing 63.10% performance on the VQA v2 val, and the addition of VGQE showing 63.18% performance on the general VQA dataset.

5.4 CSS (LM+H)

Chen et al. [25] proposed **Counterfactual Sample Synthesis** (**CSS**) to improve VQA performance, focusing on visual explainability and question sensitivity. CSS consists of two components: V-CSS, which creates counterfactual images by masking key objects, and Q-CSS, which replaces important words in questions with [MASK]. This training approach helps the model focus on relevant visual regions and words. Experiments on VQA-CP v2 and VQA v2 showed that CSS, applied to Up-Down with LMH, achieved state-of-the-art results: 58.95% on VQA-CP v2 test, 59.91% on VQA v2 val, and 60.95% on VQA-CP v1. The authors also introduced a new metric, confidence improvement, to assess question sensitivity.

5.5 CIKD (Up-Down)

Pan et al. [118] introduced **Casual Inference with Knowledge Distillation** (**CIKD**) to address the limitations of traditional VQA models [18, 34], which often fail to capture the relationship between visual information and answers due to reliance on statistical correlations. CIKD uses causal inference [120, 121] to reduce bias and improve robustness. The model constructs a causal graph and explores the causal effect between questions and answers, using counterfactuals derived from a question-only model. To improve generalizability, CIKD incorporates knowledge distillation, but it also handles biases introduced by this process using ensemble learning. Their experiments on VQA-CP v2 and VQA v2 achieved 54.05% and 62.98% performance, respectively.

5.6 WMA (MLB and MCB)

Li et al. proposed the Multi-Scale Self-Attention Convolutional Network (WMA) [86] for VQA tasks, combining self-AMs and multiscale deep networks to improve feature extraction. By using weight-sharing convolutional networks with residual modules, WMA simplifies the model while enhancing performance. Experiments on the PASCAL VOC 2012 [90] and VQA v1/v2 datasets showed that WMA improved performance by 1.09% and 1.12% over previous methods like MCB [40] and MLB [78], achieving 67.79% and 67.89% on VQA v2. The model successfully handles objects of various sizes while remaining lightweight.

5.7 GGE (Up-Down)

Han et al. [53] proposed the **Greedy Gradient Ensemble** (**GGE**) framework to tackle language bias in VQA tasks. They identified two types of biases: (1) Distribution Bias, arising from differences in training and test set distributions, and (2) Shortcut Bias, where models rely on specific QA pair patterns instead of proper visual grounding. GGE leverages overfitting to train models on ideal data distributions and address difficult problems overlooked by biased models. Using Up-Down [7] as the base, GGE-D targets distribution bias, GGE-Q addresses shortcut bias, and GGE-DQ tackles both. Experiments on VQA-CP v2 and VQA v2 showed the GGE-DQ model achieving 57.32% and 59.30% performance, respectively.

5.8 PromptCap (OFA)

Hu et al. [61] used the excellent knowledge discovery and reasoning capabilities of **large language models** (**LLMs**), such as GPT-3, to solve knowledge-based VQA problems. Because the input of the language model is text, the authors present prompt-guided image captioning (PromptCap) to generate appropriate captions for images. The captions generated by PromptCap contain more

detailed information than human annotations, which are suitable for use as inputs for VQA. In addition, the authors further improved the performance by defining a new Soft VQA Accuracy to select captions that were more relevant to the question and correct answers among the multiple captions generated using PromptCap. In the experiment, they demonstrated 60.4% accuracy for the OK-VQA dataset and 59.6% accuracy for AOKVQA. PromptCap is limited by the fact that it requires additional work to construct a new dataset by adding human-written examples to the existing VQA dataset, and it focuses only on VQA tasks in the knowledge base; however, we showed its scalability by using NLVR2 as an example.

5.9 BEiT-3 (VLMo)

Wang et al. proposed BEiT-3 [150], a multimodal model leveraging BERT Pre-Training for Image Transformers, designed for diverse downstream tasks. BEiT-3 offers key advantages: (1) it uses Multiway Transformers (VLMo) [14] to share parameters across modalities, enabling alignment and multimodal fusion; (2) it employs a single, scalable pre-training task, mask-then-predict, treating images as a "foreign language" like text; and (3) it scales up with billions of parameters and extensive training data, including 15M images, 21M image-text pairs, and 160 GB of text. BEiT-3 achieved state-of-the-art results, with 84.19% on VQA v2 test-dev and 92.58% on NLVR2 test-P, showcasing its extensibility for integrating other modalities like audio.

5.10 Prismer (Mask2Former and 5 others)

Liu et al. [97] introduced Prismer, a data- and parameter-efficient model that uses an ensemble of pre-trained domain experts, including Mask2Former [30] for semantic segmentation, UniDet for object detection, and others for tasks like text detection and depth estimation. The model pools expert knowledge effectively and compresses multimodal features using an expert resampler. In experiments with ViT [37] and RoBERTa [98], Prismer achieved 78.4% on VQA v2 testdev and 78.5% on test-std, using 120 times less data than traditional models like PaLI [29] (84.3%). Despite this, Prismer has the potential for further research in areas like multimodal in-context learning and zero-shot adaptation.

5.11 CuMo (CLIP and Mistral)

Li et al. [85] argue that recent multimodal LLMs primarily focus on increasing the quantity of text-image pair data and enhancing the performance of LLMs. However, these improvements have the disadvantage of high computational costs. To address this issue, the authors propose a method to improve model performance in the visual domain. Inspired by the **Mixture-of-Experts** (**MoE**), the authors propose CuMo. CuMo incorporates Co-upcycled Top-K sparsely-gated Mixture-of-experts blocks into both the vision encoder and the MLP connector. The authors first pre-train the MLP connector, which transforms visual tokens in CuMo architecture into word embedding space, and then use it to initialize CLIP's MoE, ensuring stability and reducing costs. Additionally, they introduce auxiliary losses to consider balanced loads among experts. In experiments, they show that the performance of VQAv2 is 80.6% and that of GQA is 63.2% when using Mistral-7B. The authors trained their model on a completely open-sourced dataset, demonstrating sufficient reproducibility, and employed a three-step training method to enhance training stability.

5.12 Florence-VL (Florence-2 and Llama-3)

Chen et al. [24] argue that multimodal LLMs have been mostly dominated by LLMs. They point out that commonly used visual encoders, such as CLIP or SigLIP, often overlook pixel or region features that are important for various downstream tasks because they provide the overall image features in the last layer. So the authors propose Florence-VL, which uses rich visual representations generated

Table 3. Overview of VQA Datasets

Data Characteristic	Dataset	Year	Image Source	# of images	# of questions	Average Question Length	Average Answer Length	OE/MC
Real-World Images Datasets	DAQUAR [104]	2014	NYU-Depth V2	1,449	12,468	11.5	1.2	OE
	VQA v1-real [8]	2015	MS-COCO	204,721	614,163	6.2	1.1	OE/MC
	FSVQA [138]	2016	MS-COCO	204,721	369,861	-	6.1	OE
	Visual Genome [80]	2017	YFCC100M, MS-COCO	108,000	1,773,258	5.7	1.8	OE
	VizWiz [51]	2018	Human	31,173	31,173	6.7	1.7	OE
	NLVR2 [142]	2018	Google Images search engine	107,292	-	14.8	1.0	MC
	OK-VQA [108]	2019	MS-COCO	14,031	14,055	8.1	1.3	OE
	VQA360° [31]	2020	Stanford 2D-3D, Matterfort3D	1,490	16,945	-	-	OE/MC
	WorldCuisines [152]	2024	Wikipedia, Wikimedia Commons	6,045	1,152,000	-	-	OE/MC
	VAGUE [113]	2024	Visual Commonsense Reasoning	3,996	3,996	16.6	-	MC
Synthetic Images Datasets	VQA v1-abstract [8]	2015	MS-COCO	50,000	150,000	6.2	1.1	OE/MC
	CLEVR [69]	2017	Synthesized Scene	100,000	999,968	18.4	-	OE
	NLVR [141]	2017	Synthetic Shapes	92,244	-	11.2	1.0	MC
	DocVQA [109]	2021	UCSF	12,767	50,000	8.1	2.2	OE
	IconQA [101]	2021	IXL-Math Learning	96,817	107,439	8.4	-	MC/Filling-in-the-blank
	SlideVQA [144]	2023	slideshare	52,480	14,484	-	-	OE
	IllusoryVQA [131]	2024	MNIST, Fashion-MNIST	27,346	27,346	-	1.0	MC
Unbiased Datasets	VQA v2 [45]	2017	MS-COCO	204,721	1,105,904	-	-	OE/MC
	VQA CP v1 [3]	2018	MS-COCO	195,000	370,000	-	-	OE/MC
	VQA CP v2 [3]	2018	MS-COCO	219,000	658,000	-	-	OE/MC
	GQA [65]	2019	Visual Genome	113,018	22,669,678	7.9	-	OE
Others	VQA-RAD [83]	2018	MedPix	315	3,515	-	1.6	OE/MC
	PathVQA [56]	2020	Textbook of Pathology, Basic Pathology, PEIR digital library	4,998	32,779	9.5	2.5	OE/MC

In this table, OE and MC stand for open-ended and multiple-choices, respectively.

by Florence-2 [156]. Here, to integrate the strengths of Florence-2 into MLLMs, the authors propose **Depth-Breadth Fusion** (**DBFusion**). To expand the breadth of visual representations, prompts for Captioning, OCR, and Grounding tasks are used to identify detailed information about various spaces required for each task. Next, for the depth of visual representations, the authors combine the lower-level features extracted by DaViT with the higher-level features extracted by the three previous prompts. To combine the features, the authors compared three fusion processes: Token Integration, Average Pooling, and Channel Integration, using the LLaVA-1.5 [96] dataset, and chose Channel Integration, which is simple but effective. In the experimental section, the authors showed that when Llama-3-8B-Instruct was used as a backbone, it achieved 84.7% performance on VQAv2 and 59.1% performance on VizWiz. The authors argue that further improvements to DBFusion could be made. One regret is that there is insufficient evidence for the selection of three tasks for Breadth.

6 Datasets and Evaluation

Early VQA datasets were based on MS-COCO [90] (328,000 images), but the noise and bias in real images limited accurate reasoning validation. To address this, synthetic datasets (e.g., CLEVR [69]) were introduced, enabling the evaluation of models' ability to understand spatial and logical relationships. However, dataset bias remained a significant issue. Unbiased datasets (e.g., VQA-v2 [45], VQA-CP [3]) with different answer distributions in training and test sets were proposed to reduce reliance on superficial correlations, although they caused performance drops in existing models. Additionally, the lack of goal-specific datasets (e.g., VizWiz [51]) was highlighted, emphasizing the need for targeted datasets to advance VQA research. Table 3 lists the VQA datasets used in this study.

6.1 Real-World Image Datasets

6.1.1 DAQUAR. DAQUAR [104] was the first benchmark for VQA tasks. The images were built using NYU-Depth V2 [139]. In NYU-Depth V2, the objects in an image are semantically segmented, and every pixel is labeled with an object class. Of the 1,449 images, 795 were classified as training images and 654 as test images. There were 12,468 question-answer pairs, with 6,794 training pairs and 5,674 test pairs. There were approximately nine questions per image. The authors also proposed a **WUP Set** (**WUPS**) score based on the WUP score [155], inspired by Fuzzy Sets, for

performance measurement. DAQUAR is a sufficiently good first benchmark for the VQA task; however, the amount of data is too small to train the model sufficiently. In addition, the images were limited to indoor images.

6.1.2 VQA v1 Real Images. VQA v1 real [8] was one of the most popular datasets used in the early days of VQA research and is still used today. VQA v1 is divided into real and abstract images, which are discussed here. The images were built using the COCO dataset, which used 123,287 images for training and 81,434 images for testing. The questions were generated by human annotation and varied, and the authors provided the subjects with the previous questions that had already been asked. The open-ended questions allowed for a variety of answers, but most were yes or no. There were 10 Q&A pairs for each image. However, the VQA v1 dataset has a language bias in which the questions and answers are sufficient to be answered without looking at the image. The VQA v2 evolved into a dataset that addresses this problem.

6.1.3 FSVQA. The **Full-Sentence Visual Question Answering** (**FSVQA**) [138] was proposed by Shin et al. to address the limitations of existing VQA [8] datasets, where the correct answers are mostly composed of one word and do not have the richness of realistic human answers. Just as the CV field is evolving from simply detecting objects in images to generating sentences that describe images by utilizing information from the images, VQA will also evolve from simply generating single-word answers to generating answers in the form of natural sentences. To reduce the cost of generating answers one by one, the authors applied linguistic rules to the short answers in the existing VQA [8] datasets and converted them into complete sentences.

6.1.4 Visual Genome. Visual Genome [80] was proposed by Krishna et al. in response to the fact that there are many benchmarks for perceptual tasks such as classical image categorization, but there are no benchmarks for cognitive tasks that require understanding the interaction between objects, such as image description or question and answer.

6.1.5 VizWiz. The VizWiz dataset [51] is the first of its kind for visually impaired people. The images were captured by blind people on their mobile phones, and the questions were based on spoken questions regarding the images. This study builds on a previous study [15] that collected 72,205 visual questions over four years. However, many of these images contain personal information such as card numbers, personal emails, and account numbers. Additionally, some images were too blurry or bright to be identified. Therefore, the authors did a lot of filtering and ended up with 31,173 pairs of visual questions. Most images are poorly lit and out of focus; therefore, one needs to be able to answer them appropriately, even with low-quality images. This dataset has a clear purpose and is more challenging than a typical VQA dataset; however, it is interesting. This dataset has the potential to advance assistive technologies for visually impaired populations.

6.1.6 NLVR2. NLVR2 [142], an extension of NLVR [141], addresses the limitations of synthetic images and languages used in previous datasets like Cornell NLVR. It involves real images retrieved via Google Image Search and human-written descriptions, offering 107,292 image-sentence pairs. This dataset improves semantic diversity and realism, making it more applicable to real-world VQA tasks involving reasoning about quantities, comparisons, and relationships.

6.1.7 OK-VQA. Marino et al. [108] introduced OK-VQA, a dataset designed to produce more human-like answers in VQA. Unlike traditional datasets that focus on simple questions (e.g., counting, color judgment), OK-VQA requires external knowledge to answer questions, making it more complex. It tests the ability to use knowledge from the web, databases, and other sources unrelated to the image itself. While OK-VQA contains 10 types of knowledge, the best-performing model on

this dataset achieved 66.8% accuracy, significantly lower than the 84.3% accuracy of the top model on the VQA v2 benchmark.

6.1.8 VQA 360°. Currently, most VQA datasets are dominated by images with a normal field of view. However, using images from a 360° camera allows us to utilize all information from the front, sides, top, bottom, and back. It is possible for future robots to use 360° cameras. Although there is currently a wide range of studies utilizing 360° images [59], it is significant that the authors proposed the first benchmark to provide Q&A with 360° images [31]. The images focused on indoor scenes for sophisticated reasoning and were based on Stanford 2D-3D and Matterport 3D. The image size was 1024 × 512 pixels. The questions were divided into five categories based on the template. The total number of QA pairs was 16,945 with approximately 11 questions per image. VQA 360° images have the disadvantage of being more distorted than flat images owing to their panoramic nature.

6.1.9 WorldCuisines. **Vision Language Models** (**VLMs**) have mainly struggled to handle non-English languages and poorly understood cultural contexts, so Winata et al. propose WorldCuisines [152], which has 1,152,000 text-image pairs for 30 languages and dialects across 9 language families, to evaluate and improve multilingual and multicultural performance. It is the largest multicultural VQA benchmark. Food also serves as a proxy for cultural knowledge and plays an important role in shaping language [39]. The authors split the benchmark into two tasks: predicting the name of the food and location prediction, which is to guess which region the food is in. The images are taken from Wikipedia, with 6,045 images for 189 countries. The QA pairs for this dataset are 1,080,000 train set, 12,000 test small set, and 60,000 test large set. This dataset has QA pairs that are not comparable to the existing dataset for multicultural VQA, but it uses images from the United States at a rate of 9.47%, which is the highest rate among 6,045 food images for 189 countries, so it seems that the ratio needs to be adjusted.

6.1.10 VAGUE. The authors argue that the questions in most current VQA datasets are simple and direct, overlooking the ambiguity that can occur in human communication. Similar studies have used FLUTE [21] and SBU [82] as data for finetuning T5, but both datasets were limited to single-modality training. Nam et al. propose VAGUE: Visual Contexts Clarify Ambiguous Expressions [113] to improve disambiguation in multimodal contexts. The images are sourced from the **Visual Commonsense Reasoning** (**VCR**) dataset and filtered based on the authors' "interestingness" criterion, which prioritizes ambiguous and contextually rich scenarios. The VAGUE benchmark requires models to select the intended meaning of an ambiguous utterance from four choices, using the multimodal context provided by the image. The dataset consists of 3,996 images, with object recognition facilitated by the bounding boxes from VCR and the **Recognize Anything Model** (**RAM**). While the dataset construction is sophisticated and represents a novel approach to enhancing disambiguation in multimodal VQA, the relatively small size of 3,996 samples may limit its generalizability and robustness for large-scale applications.

6.2 Synthetic Image Datasets

6.2.1 VQA v1 Abstract Scenes. The VQA v1 abstract [8] removed the complexity and noise of VQA v1 real images to enable high-level image inference. It also included five captions per image. The abstract image contained 20 "paperdoll" human models with eight different facial expressions based on sex, race, and age.

6.2.2 CLEVR. The CLEVR dataset [69] was designed to test the reasoning abilities of VQA models, addressing biases in traditional datasets. CLEVR uses 3D shapes and simplified compositions to focus on reasoning, where answers can only be derived from the visual information in the image,

avoiding surface-level shortcuts. It contains 100,000 images and 999,968 automatically generated questions across 90 categories. While effective for testing reasoning, CLEVR's lack of generic images makes it less suitable for general VQA tasks.

6.2.3 NLVR. Cornell **Natural Language Visual Reasoning** (**NLVR**) [141] is a benchmark that, given a synthesized image with several shapes and a sentence given as a description of the image, selects true/false whether the sentence correctly describes the image. The image was generated by random rendering, and the description of the image was written according to a template using a crowdsourcing platform. Unlike the visual reasoning task CLEVR [69], which involves synthesized language, NLVR uses human-generated sentences.

6.2.4 DocVQA. Although **Document Analysis and Recognition** (**DAR**) have been extensively studied, research has tended to be module-specific rather than integrated reasoning, such as table extraction [74] or character recognition [35] from images. Therefore, Mathew et al. proposed a higher-level task called **Document Visual Question Answering** (**DocVQA**) [109] that conditionally interprets images. To answer these questions, DocVQA needs to collect and utilize not only the text in the image, but also information such as tables and checkboxes, which are components of the document. The document Images were based on the UCSF Industry Documents Library, from which the authors handpicked images to reduce the bottleneck in VQA. In addition, the QA pairs were created by humans through a three-step process. The significance of DocVQA is that it provides a benchmark in the field of DAR that allows researchers to move from modular to integrated reasoning.

6.2.5 IconQA. Lu et al. proposed the **Icon Question Answering** (**IconQA**) dataset [101], consisting of 107,439 Q&A pairs across three tasks: multiple-image-choice, multiple-choice, and filling-in-the-blank. Along with IconQA, they introduced Icon645, containing 645,687 color icons across 377 object classes for pre-training. The data, based on IXL Math Learning, are designed for mathematical word problems targeting children, focusing on skills like geometry, counting, and comparison across 13 categories. Unlike other datasets, IconQA was tested for data bias, showing no bias in a blind study, confirming its reliability.

6.2.6 SlideVQA. Tanaka et al. [144] introduced SlideVQA, a dataset focused on inferring answers across multiple images from slide decks. It includes 2,619 slide decks with 52,480 slides on 39 topics, offering 14,484 QA pairs–12,466 for single-hop and 2,018 for multi-hop. The dataset is based on SPaSe [54], with questions designed to require referencing all images in a deck. While it advances multi-image VQA, the dataset's single-hop questions still dominate, highlighting the need for more multi-hop data.

6.2.7 IllusoryVQA. Illusory VQA [131] is a dataset proposed to enable a multimodal model to recognize optical illusions that may occur in images like humans. The dataset consists of IllusionMNIST, IllusionFashionMNIST, IllusionAnimals, and IllusionChar. It uses image sources from MNIST and Fashion-MNIST. For IllusionAnimals, images for 10 animals were generated using SDXL-Lightning [89], and the authors created the images for IllusionChar. With the image sources, ControlNet [167] was used to generate optical illusion images, and the Description for generating optical illusion images was used with ChatGPT, Gemini, Mixtral-8x7B-Instruct-v0.1, and Gemma-1.1-7b-it to generate 1027 descriptions that ensure diversity. As a result, a total of 27,346 samples were generated, consisting of 20,460 training sets and 6,886 test sets. The authors used human validation and evaluation when creating the dataset, and used NSFW (Not Safe for Work) detector models to filter out potentially offensive images. However, the limitation of this dataset is that there is only one object in the image that creates an optical illusion.

6.3 Unbiased Datasets

6.3.1 VQA v2. The VQA v2 dataset [45] is more balanced than the original VQA v1 dataset, with twice the number of questions per image. The reason for this balance is that the ideal solution to the VQA task is to answer questions based on both vision and questions. However, existing models ignore visual information and exploit certain patterns between questions and answers to improve performance. The images were the same as in VQA v1 (123,287 for training and 81,434 for testing), but the questions were twice as many as in VQA v1 (658,111 for training and 447,793 for testing). There were ten correct answers for each question. VQA v2 reduces the bias in the dataset and helps measure the correct performance of the VQA model.

6.3.2 VQA-CP. The VQA under Changing Priors (VQA-CP v1, VQA-CP v2) [3] dataset differs from the original VQA v1 and VQA v2 datasets by changing the distribution of answers to questions in the training and test sets. The reason for the different distributions is to eliminate the language bias of ignoring visual information. If the answer distributions of the training and test are the same, there will be language bias in inferring the correct answer for a particular question without considering visual information. To do this, the authors first group questions with the same question type and answer, and then greedily redistribute them so that the distributions of the training and test sets are not the same.

6.3.3 GQA. GQA is a dataset that uses real-world images to address the problems with the current VQA dataset. Current VQA datasets contain biases. For example, most bananas are yellow. This statistical bias and skewed distribution of answers cause the model to not focus on visual information. This problem is exacerbated by the simplicity of the questions, as most of them rarely require anything beyond object recognition. In addition, the lack of content annotation makes it difficult to determine what is affecting performance. To address this, the authors leverage the Visual Genome scene graph [80] and a newly created linguistic grammar to generate questions with visually grounded and structured representations in image scene graphs. GQA is inspired by CLEVR, but CLEVR is not an effective dataset due to its limited diversity. In contrast, GQA is more diverse as it incorporates real images. GQA consists of 113,018 images and 22,669,678 questions. The question engine controlled the distribution of answers to remove bias when generating questions, and used a new adjustable smoothing technique to mitigate question bias. The questions use 250 manually configured patterns and 274 patterns derived from VQA v1. Questions generated using patterns may have high diversity but may not be natural. The author also presented five evaluation metrics: Consistency, Validity, Plausibility, Distribution, and Grounding to check the ability of the inference model.

6.4 Others

6.4.1 VQA-RAD. VQA-RAD [83] is a medical imaging VQA dataset that provides medical professionals with a reference to natural questions and answers regarding radiology images. The image source was constructed using the open radiology archive MedPix. The medical image data comprised 104 head axial single-slice CT or MRI images, 107 chest X-rays, and 104 abdominal axial CT images. The questions were guided by those that clinicians could ask their peers or radiologists, and all question-answer pairs were manually validated and categorized.

6.4.2 PathVQA. PathVQA [56] is a pathology dataset that was proposed to develop an "AI Pathologist" capable of passing the American Board of Pathology examinations. This is the first dataset of pathological VQA. The image sources are two books (Textbook of Pathology and Basic Pathology) and the **Pathology Education Informational Resource** (**PEIR**) digital library, which generate question-answer pairs and collect images. The dataset was built by developing

a semi-automated pipeline to extract pathology images and captions from books, and generate question–answer pairs from the captions using NLP.

6.5 Evaluation Metrics

Evaluating the natural language generated by a VQA model is a complex task. For a correct evaluation, it is necessary to evaluate sentences grammatically and semantically. Questions in the VQA tasks can be broadly categorized into open-ended and multiple-choice tasks. Open-ended questions are more challenging than multiple-choice questions, in which the user must choose between several options because sentences are generated without any structured format. In this section, we introduce metrics for evaluating natural languages generated by VQA models. Until now, metrics that examine only simple answers to simple questions in VQA tasks have been mainstream; however, considering that future VQA research will require sufficiently long answers to complex and abstract images and questions, we also introduce a sentence-based evaluation metric.

6.5.1 VQA Accuracy. An appropriate assessment of the open-ended questions was proposed in the VQA [8]. Ep. 1 is an expression that is accepted as the correct answer if at least three of the multiple ground truths given for the question are the same. For the VQA [8] dataset, approximately 10 ground truths were provided for each question, which can be described as follows:

$$Accuracy_{VQA} = \min\left(\frac{A\,specific\,answer\,is\,given\,by\,\#\,subjects}{3}, 1\right). \tag{1}$$

However, this equation has problems that must be addressed. However, this equation is expensive because it requires multiple ground truths. In addition, for yes/no questions, there is a probability that the ground truth contains at least three answers; therefore, any answer generated has a chance of being correct. Finally, the generated natural language must match exactly the three ground truths to be treated as a correct answer. However, when evaluating the answers generated by questions that start with "why", the probability of a match is significantly lower.

6.5.2 BLEU. BLEU [119] score is the degree to which the words in the generated sentence are included in the ground-truth sentence. It is expressed as

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(l-r/c)} & \text{if } c \leq r \end{cases}, \tag{2}$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{3}$$

where p_n represents the precision for each n-gram and w_n is the weight for each n-gram, which adds up to 1. BP is the brevity penalty, which is an overfitting correction for the sentence length. This evaluation method is widely used in machine translation and was first applied to VQA tasks using VizWiz [51]. Currently, it is widely used in medically related VQA. However, because BLEU does not evaluate the semantic score between words, it judges the use of different words with the same meaning as incorrect. In addition, because BLEU does not assign a weight to each word, it equates the incorrect use of important words in a sentence with the incorrect use of unimportant words.

6.5.3 ROUGE. **Recall-Oriented Understudy for Gisting Evaluation** (**ROUGE**) [88] is an evaluation method based on n-gram recall. This is slightly different from BLEU, which is based on

n-gram precision. The equation used is as follows:

$$ROUGE - N = \frac{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count(gram_n)}. \quad (4)$$

In Equation 4, *RS* stands for ReferenceSummaries. This evaluation method is primarily used for text summarization and machine translation. ROUGE-N can obtain the Precision and Recall for each n-gram. Recall is the number of words in the ground truth that overlap with words in the generated answer divided by the number of words in the ground truth. However, if the model-generated answer is long, it may have a high recall score, even if the answer is not related to the ground truth. To solve this problem, we calculated precision. Precision is the number of words in the generated answer that overlap with the words in the ground truth, divided by the number of words in the generated answer. In Ref. [88], the authors presented various evaluation methods, such as ROUGE-N, ROUGE-L, ROUGE-S, and ROUGE-SU; however, ROUGE is also limited because it does not evaluate words semantically.

6.5.4 CIDEr. The sentence similarity test used by the CIDEr [148] captures both the grammaticality and the precision and recall of the sentences being compared. This is based on the concept of TF-IDF. Based on this approach, the goal of the CIDEr evaluation method is to measure the sentence similarity or consensus with most ground truths. The formula used is as follows:

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|}. \quad (5)$$

Let $g^n(c_i)$ and $g^n(s_{ij})$ be the vectorization of the TF-IDF values for each n-gram, where s_{ij} is the sentence in the ground truth and c_i is the sentence generated by the model. Next, we compute the cosine similarity of $g^n(c_i)$ and $g^n(s_{ij})$, that is, we find the TF-IDF vectors in the model-generated sentences and ground truth sentences and measure the cosine similarity of the two vectors. The final CIDEr is expressed as follows:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i). \quad (6)$$

Where N is the value of n in the n-gram. The authors empirically found that N works best when it is four. However, in the case of CIDEr, accurate performance measurement requires a large number of ground truth sentences to be compared with the model-generated sentences.

6.6 Evaluation

In this section, we present the performance of VQA v2 [45], which is traditionally used to evaluate the performance of VQA models, and VQA-CP v2 [3], which changed the distribution of correct answers to questions, in Table 4. In addition, we present the performance of GQA [65], OK-VQA [108], DocVQA [109], CLEVR [69], NLVR2 [142], and A-OKVQA [132] in Table 5. We also provide an analysis of the performance of each benchmark.

6.6.1 Performance Analysis.

— **VQA v2:** VQA v2 [45] has been a foundational benchmark for VQA research since the early models and remains widely used today. VLMo [14], which utilizes the AM, achieved a high score of 82.88%. BEiT-3 [150] improves upon VLMo by approximately 2% using the Masked Data Modeling technique introduced during pretraining and by scaling up the pretraining data. Additionally, BEiT-3 can handle multiple tasks, including VQA, image captioning, and object detection. Later models, such as CuMo [85] and Florence-VL [24], are multimodal LLMs that integrate LLMs with VLMs rather than being specifically designed for VQA tasks.

Table 4. Performance of Some of the VQA Models Covered in This Survey on VQA v2 [45] and VQA-CP v2 [3]

Model	Year	VQA v2 test-dev				VQA v2 val	VQA v2 test-std	VQA-CP v2 test
		All	Y/N	Num.	Other	All	All	All
BAN [77]	2018	70.04	85.42	54.04	60.52	-	70.35	-
Up-Down [7]	2018	-	-	-	-	63.20	70.34	-
RUBi (MuRel) [18]	2019	63.18	-	-	-	61.16	-	47.11
Learned-Mixin (Up-down) [34]	2019	-	-	-	-	-	-	48.78
Learned-Mixin+H (Up-down) [34]	2019	-	-	-	-	56.35	-	52.01
SelRes [58]	2020	72.70	-	-	-	-	72.80	-
VGQE (MuRel) [81]	2020	-	-	-	-	65.73	-	50.11
CSS (LM+H) [25]	2020	-	-	-	-	59.91	-	**58.95**
MDFNet [169]	2021	71.19	86.85	53.73	61.78	-	71.32	-
N-KBSN [103]	2021	71.14	87.13	54.05	60.9	-	71.23	-
UFSCAN [168]	2021	70.46	85.52	**54.99**	61.08	-	70.73	-
CIKD (Up-Down) [118]	2021	-	-	-	-	62.98	-	54.05
WMA (MLB and MCB) [86]	2021	67.89	85.15	39.85	58.32	-	-	-
GGE (Up-Down) [53]	2021	-	-	-	-	59.30	-	57.32
DMBA-NET [158]	2022	70.69	87.55	51.15	60.72	69.45	70.85	-
LSAT [136]	2022	71.67	**87.74**	54.51	**61.83**	-	71.94	-
VLMo [14]	2022	82.88	-	-	-	-	82.78	-
PromptCap (OFA) [61]	2022	-	-	-	-	**74.10**	-	-
BEiT-3 (VLMo) [150]	2023	**84.19**	-	-		-	**84.03**	-
CSCA [110]	2023	70.72	86.57	53.58	61.06	67.36	71.04	-
Prismer (Mask2Former and 5 others) [97]	2023	78.43	-	-	-	-	78.49	-
ACOCAD [9]	2024	71.02	87.43	53.67	60.90	-	71.18	-
CuMo (CLIP and Mistral) [85]	2024	82.20	-	-	-	-	-	-
Florence-VL (Florence-2 and Llama-3) [24]	2024	82.10	-	-	-	-	-	-

Results in bold are the best performers on each benchmark.

Table 5. Performance of Some Models on GQA [65], OK-VQA [108], DocVQA [109], CLEVR [69], NLVR2 [142], and A-OKVQA [132]

Dataset	Model	Year	Accuracy	Dataset	Model	Year	Accuracy	Dataset	Model	Year	Accuracy	Dataset	Model	Year	Accuracy	Dataset	Model	Year	Accuracy
GQA [65]	Up-Down [7]	2018	49.74	OK-VQA [108]	PICa [159]	2021	48.00	DocVQA [109]	LayoutLMv2 [157]	2021	78.08	CLEVR [69]	FiLM [122]	2017	97.70	NLVR2 [142]	VLMo [14]	2021	86.86
	NSM [63]	2019	62.95		Flamingo 80B [5]	2022	50.60		MatCha [95]	2022	74.20		NS-VQA [161]	2018	**99.80**		BLIP [84]	2022	83.09
	GRN [48]	2019	61.22		REVIVE [91]	2022	56.60		UDOP [145]	2023	84.70		MAC [64]	2018	98.90		CoCa [162]	2022	87.00
	LXMERT [143]	2019	60.00		PNP-VQA [147]	2022	35.90		DUBLIN [1]	2023	78.20		DDRprog [140]	2018	98.30		BEiT-3 [150]	2022	**92.58**
	CFR [116]	2021	**72.10**		PromptCap [61]	2022	60.40		Qwen-VL-Plus [12]	2023	90.24		QGHC [42]	2019	86.30		X-FM [170]	2023	88.40
	VinVL+L [149]	2023	64.85		PaLI-X [27]	2023	66.10		PaLI-3 (w/ OCR) [28]	2023	88.60		NS-CL [106]	2019	98.90		LXMERT [143]	2019	25.90
	ACOCAD [9]	2024	57.37		Lyrics [99]	2023	58.20		Omni-SMoL [153]	2024	90.60		LCGN [60]	2019	97.90		PromptCap [61]	2022	59.60
	CuMo [85]	2024	64.90		Prophet [163]	2023	62.50		Florence-VL [24]	2024	82.10		MDETR [73]	2021	99.70	A-OKVQA [132]	GPV-2 [72]	2022	40.70
	Lyrics [99]	2024	62.40		PaLI-X-VPD [62]	2024	**66.80**		ScreenAI 5B [10]	2024	89.88		OCCAM [151]	2021	99.40		Prophet [163]	2023	58.50
	Florence-VL [24]	2024	61.80		HYDRA [75]	2024	48.60		MLCD [6]	2024	**91.60**		NeSyCoCo [71]	2024	99.70		PaLI-X-VPD [62]	2024	**68.20**

Results in bold are the best performers on each benchmark.

These models have demonstrated performance comparable to VQA models in zero-shot settings. This suggests that future advancements may shift toward multimodal LLMs with fine-tuning strategies specifically for VQA.

— **GQA:** GQA [65] is a representative benchmark for solving the problem that biased answer distribution causes the model to ignore the relationship between images and questions. This is a very difficult problem and it is difficult to improve performance unless modeling focuses on this problem. Here, multimodal LLMs such as CuMo [85] and Florence-VL [24] still follow the existing performance well, but CFR [116], which is proposed to reduce the semantic gap between images and questions without learning biased answer distributions and taking shortcuts, showed the best performance. In the future, in the development of multimodal LLMs with fine-tuning strategies for VQA, it will be important to have a strategy to not ignore the relationship between images and questions by learning biased answer distributions.

— **OK-VQA:** OK-VQA [108] can evaluate whether the model has the ability to provide rich answers by referencing external resources in addition to the information in the image, rather

than a simple yes or no answer. There have been attempts to improve the performance of knowledge-based VQA by intentionally focusing on regional visual representations, such as REVIVE [91], and utilizing knowledge retrieval for all objects in the image. However, since 2022, research using LLMs, such as PromptCap [61], has been dominating performance on knowledge-based VQA, using the enormous size of LLMs and the knowledge accumulated from huge training data.

— **DocVQA:** DocVQA [109] is a dataset that has led research on documents toward integrated reasoning rather than module-specific research such as DAR. Most of the models that perform well are multimodal LLM models, but MLCD [6] is the best performer. There are many objects in the document that form a semantic structure. Cluster discrimination was proposed to encode this semantic structure, but it only provides a single pseudo-label for the image, so MLCD [6] was proposed to receive the multi-label signal. This means that multimodal LLM still neglects the multi-label signal.

— **CLEVR:** CLEVR [69] evaluates a model's performance in attribute identification, counting, comparison, spatial relationships, and logical operations using images consisting of 3D shapes. It has minimal bias towards the correct answer. It also provides a very intuitive yet simple way to evaluate whether a model makes sophisticated inferences. While most of them perform close to 99%, QGHC [42] does not. This is because it initially uses a Question-guided Hybrid Convolution strategy for strong interaction between text and visual information, which allows for a close coupling of text and visual information, but also suggests that combining low-dimensional feature spaces can be limiting for learning complex relationships when multi-level inference is required, such as in CLEVR.

— **NLVR2:** NLVR2 [142], which determines whether a description is true based on information about an image and a given description of the image, can be better applied to real-life environments because it does not use synthetic images or synthetic languages. This benchmark reveals the difference in performance between a model that focuses on the AM and a model that focuses on model agnosticity. The performance of a model that focuses on the AM, such as XFM [170], which improves the structure by configuring a language encoder, a vision encoder, and a fusion encoder and suggests a new learning method, did not exceed 90%. However, when focusing on model agnosticity, such as BEiT-3 [150], which adopted VLMo [14] as it is and improved the external part of the model, such as using the Masked Data Modeling technique proposed by the authors in the pretraining process, it showed a performance of 92.58%.

— **A-OKVQA:** A-OKVQA [132] is an Augmented OK-VQA dataset, requiring more extensive knowledge than OK-VQA. LXMERT [143], which utilizes the existing AM, performed poorly with an accuracy of 25.9%. GPV-2 [72] also relies on AMs. After extensive training on five datasets, including OpenImages, VisualGenome repositories, and a web-derived dataset with over 10k concepts, GPV-2 achieved an improved accuracy of 40.7%. However, multimodal LLMs such as PromptCap [61], Prophet [163], and PaLI-X-VPD [62] demonstrate superior performance, achieving approximately 60%. These results highlight the effectiveness of leveraging LLM knowledge for VQA tasks, especially when external knowledge is required, rather than modifying model structures or relying on extensive training datasets.

7 Future Research

7.1 Future Research Directions

Research on VQA has grown rapidly since it was first presented in Ref. [8] in 2015. In addition, high-performing models have been proposed in the fields of vision and language, and good performances have been demonstrated by importing and applying them to VQA. However, VQA

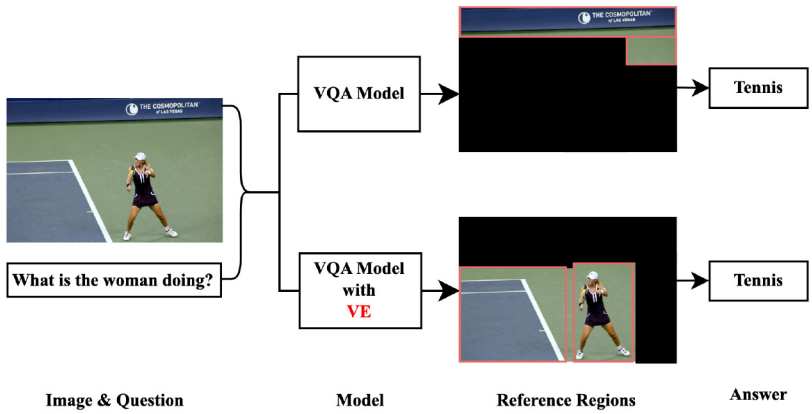


Fig. 5. Examples of reference regions for models with and without visual explainability (VE = Visual Explainability).

considers vision and language equally important and uses each piece of information appropriately and correctly. From this perspective, JE and AMs have been studied in the field of VQA research to properly fuse visual and language information, and MA methods have been studied to overcome the limitations caused by datasets and model learning. An ideal VQA model would utilize core information to generate an answer from the given visual information and question and, in exceptional circumstances, utilize external knowledge that is not present in the given information to generate an answer. In this section, we present the current limitations of VQA tasks and directions for overcoming these limitations.

— **Bias of dataset:** Biases are elements that cannot be eliminated from the dataset. This is sometimes referred to as **Long-Tail Distribution (LTD)**. For example, in a dataset, the correct answer to the question "What color are the bananas in the image?" will be mostly yellow. However, there may be images of unripe bananas, and an ideal VQA model would recognize that most bananas are yellow, and would be able to cope well with the exceptional situation of unripe bananas. However, most VQA models do not make proper use of visual information owing to biases in the dataset, and rely on question-and-answer distribution to determine the correct answer. MA methods or models have been proposed to solve this problem since the early days of VQA research, and benchmarks such as VQA-CP and other recently proposed datasets have strived to have minimal bias. However, there is a significant lack of benchmarks and models specializing in bias. In Ref. [53], bias was analyzed by dividing it into distribution and shortcut biases, and methods for removing each bias were proposed. In addition, in Ref. [25], counterfactual data augmentation was proposed to eliminate bias. In future, a more detailed analysis and research to remove bias will be needed.

— **Visual-explainability:** Visual-explainability is the capability of an ideal VQA model. The model uses correct visual information to make decisions [130]. For example, in Figure 5, the question "What is the woman in the image doing?" should be answered with correct information, such as what she wearing or what she is holding in her hand. This has been studied with regard to the AM [7, 94, 124], strong biases remain in the network and have not been completely solved. In addition, there has been a study that utilizes Grad-CAM [133] to calculate the contribution between objects and match human scores, but it is costly because it requires human annotation. Visual explainability is being studied not only in images but also in video question answering [87].

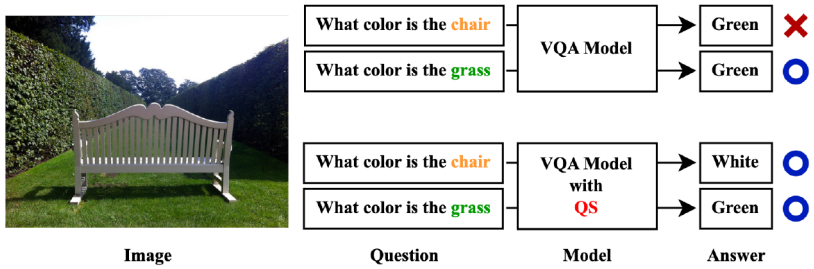


Fig. 6. Example output of a model with and without question sensitivity (QS = Question Sensitivity).

— **Question-Sensitivity:** Question-Sensitivity is one of the characteristic of an ideal VQA model. If a model understands a question well, it should be sensitive to variations. For example, in Figure 6, when we replace only one word, chair "or" grass, "in the same question format, the model recognizes that the intent of the question is different and outputs a different answer. However, most models cannot detect subtle variations. There is a study on this topic [25], and it can be seen that for an ideal VQA model, most are only focused on improving visual explainability, and there is a relative lack of research on question sensitivity. There are various models [100] that approach the AM; however, they are limited in that they cannot model the interaction between each image region and word-level text segments. Studies have been conducted [77, 117] to solve this problem, but they are limited by the lack of self-attention within this modality. Question sensitivity is the most important ability to identify and improve the desired answer. For this purpose, studies such as [25], which approached it as a framework, and [127], which proposed a multimodal fusion module, are being conducted.

— **Utilization of External Knowledge:** External Knowledge is needed when an image and a question are given, and it is difficult to answer the question using only the information given. The VQA datasets contain visual information, questions, and the answers to these questions. However, most of the questions were formatted monotonously, and the answers were short (one or two words). A model can be trained on the questions and answers in the dataset but may struggle to answer more complex questions or questions that require additional information. Therefore, to create an ideal VQA model, it is recommended to make good use of external knowledge that is not limited to the dataset. This has led to research on the use of external knowledge, such as Wikipedia [44, 107]. Recently, inspired by the PICa [159], research on the use of GPT-3 [17] has been conducted. Research on utilizing the excellent knowledge retrieval and reasoning capabilities of LLMs such as GPT-3 is being conducted in various ways [17, 47, 61]. Research on prompting LLMs such as GPT-3 or T5 [33] has been attempted in various ways other than VQA and can be meaningfully used for VQA as well.

— **Non-English VQA:** Currently, VQA research is dominated by English. This is primarily due to the fact that MS-COCO and Wikipedia, which are suitable sources for VQA benchmarks, are both dominated by English, highlighting the lack of multilingual resources. However, recent benchmark proposals for non-English VQA have emerged. As the first study, Pfeiffer et al. proposed xGQA [123], which extends the English GQA dataset to seven languages, to improve the multilingual ability of the model. However, in xGQA, the questions are provided in multiple languages, but the answers are limited to English. Later, Changpinyo et al. proposed MaXM [22], a VQA benchmark in seven languages. It uses Crossmodal-3600 [146], a multilingual image caption dataset. MaXM improves upon xGQA by supporting multilingualism in both questions and answers. It also contains a total of 2,142 questions, evenly

Table 6. Applications of Visual Question Answering

Application	Advance direction	Key element
AI-based medical image analysis assistant	Data augmentation and denoising	Explainability
Support application for the visually impaired	Protecting privacy and dealing with low-quality data	Reasoning
Video surveillance systems	Detect and track objects and analyze how they relate to each other	Real-time situational awareness
Advertising guide	A judge of whether an ad is compelling	Ability to predict and analyze sentiment

distributed across seven languages. Winata et al. [152] proposed WorldCuisines, a food-related VQA dataset across 30 languages and dialects. WorldCuisines illustrates how the same food may have different names in different countries and cultures. This suggests that research on non-English VQA needs to take a contextual approach that incorporates culture, not just language. Although various attempts have been made, most methods still rely on translating English into other languages. Furthermore, WorldCuisines, which uses the Wikipedia dataset as a source, is biased toward the United States, with the U.S. accounting for 9.47% of the 189 countries with food photos. For future research on non-English VQA, it is necessary to utilize non-English resources and create a dataset that is not biased toward English-speaking countries. In addition, standardization based on each language and cultural context will help drive interest in non-English VQA research.

7.2 Applications

The VQA has numerous real-life applications. In this section, we describe four representative areas in which VQA was applied and suggest directions for further development in each area. We also suggest the factors that are important for each application. Table 6 summarizes the areas of application.

- **AI-based medical image analysis assistant:** The first area that has been actively used is AI-based medical image analysis assistant applications. The medical domain has been actively researched, and this study was expanded by covering the medical domain VQA task in ImageCLEF 2018 [67]. Since then, the VQA-RAD [83] and PathVQA [56] datasets have been introduced, and various other medical-related VQA datasets [79, 93] have been provided to improve the performance of VQA models in medical images, which are more difficult to solve than general images. With the emergence of various datasets, various models have been developed to address the requirements of medical VQA [49]. However, data limited to the medical field are insufficient to train current models, and studies have been proposed to solve this problem [115]. A key element of VQA as a medical assistant system is its ability to sufficiently explain the basis for making such judgments.
- **Support application for the visually impaired:** The second application is a visual assistance application for the visually impaired. The reason why VQA can be used as a visual assistant application is that blind people have historically relied on volunteers or guide dogs to make judgments about visual information, and VQA can do this for them. In addition, the performance of VQA for blind people has been improved by the availability of a dataset [50, 51] in which blind people themselves take pictures and record voice questions. From the dataset, we can see that an important direction of development for visual assistance applications is the need to denoise shaky and inaccurate data, protect privacy that may be exposed, and provide appropriate reasoning on data that may not be perfect, which is a key element for VQA techniques to consider.

— **Video surveillance systems:** The third category includes video surveillance systems [32]. If VQA is used in video surveillance systems, administrators do not have to monitor and judge a large number of videos in real time, and if the ideal VQA technology is applied, they can make fairer judgments. To improve the performance of video surveillance, the ability to detect and track objects should be the first step, and the relevance of the objects to each other and to the question should be considered in situational judgment. The key element that a VQA should consider in video surveillance is its ability to make real-time situational judgments.
— **Advertising guide:** Last but not least is the advertising guide. In advertisements, the relationship between the objects in an image and sentences describing the image must be easy to understand and attractive. Thus, the first thing that VQA should do to understand advertisements is to learn what strategies the field of advertising is using and which strategies are effective. The dataset for this is [66], from which the model learns consumers' evaluations of different advertisements. The way forward for VQA in the advertising field is to be able to detect the objects present in an advertisement image and to be able to understand the correlation between the objects well, and to be able to embed the advertisement image and the sentence describing the advertisement image well. The key element that VQA should consider in the advertising field is the ability to predict and analyze what emotions a pair of advertisement images and the sentence describing the image can give to the people who encounter the advertisement.

8 Conclusion

Humans have many senses; we hear sounds, see objects, taste food, and smell. Artificial intelligence must be able to understand these different senses in order to be on the same level as humans. Among the multimodal AI fields that use these various modalities, the vision language field has received considerable attention from the visual speech recognition field to various research fields, such as image-to-text, text-to-image, VQA, and image captioning. In this survey, we summarized the background knowledge of VQA, one of the tasks in the vision language field; classified VQA models into three types: JE, AM, and MAs; and analyzed the advantages, disadvantages, and limitations of each method. We also discussed the development of datasets from the early days of VQA to the present for improved verification by dividing them into real, synthetic, and unbiased images. We also discuss the limitations of the VQA tasks and future research directions based on recent studies. We also describe a sentence-based evaluation metric for evaluating rich answers in future VQA tasks. This survey provides background on VQA research and information on where VQA research has been conducted and where it is headed.

References

[1] Kriti Aggarwal, Aditi Khandelwal, Kumar Tanmay, Owais Khan Mohammed, Qiang Liu, Monojit Choudhury, Hardik Hansrajbhai Chauhan, Subhojit Som, Vishrav Chaudhary, and Saurabh Tiwary. 2023. DUBLIN: Visual document understanding by language-image network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.emnlp-industry.65

[2] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas. 1955–1960. https://doi.org/10.18653/v1/D16-1203

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*.

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

[5] 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 *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[6] Xiang An, Kaicheng Yang, Xiangzi Dai, Ziyong Feng, and Jiankang Deng. 2025. Multi-label cluster discrimination for visual representation learning. In *European Conference on Computer Vision*. Springer, 428–444.

[7] 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 *IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV'15)*.

[9] Hesam Shokri Asri and Reza Safabakhsh. 2024. Advanced visual and textual co-context aware attention network with dependent multimodal fusion block for visual question answering. *Multimedia Tools and Applications* 83, 1 (2024), 1–28.

[10] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. 2024. ScreenAI: A vision-language model for ui and infographics understanding. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*. International Joint Conferences on Artificial Intelligence Organization, 3056–3062. https://doi.org/10.24963/ijcai.2024/339

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473. Retrieved from https://arxiv.org/abs/1409.0473

[12] 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:2308.12966. Retrieved from https://arxiv.org/abs/2308.12966

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[14] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems* 35 (2022), 32897–32912.

[15] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. VizWiz: Nearly real-time answers to visual questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST'10)*. 333–342.

[16] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *IEEE/CVF International Conference on Computer Vision*. 4291–4301.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[18] Remi Cadene, Corentin Dancette, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems* 32 (2019), 841–852.

[19] Linqin Cai, Haodu Fang, and Zhiqing Li. 2023. Pre-trained multilevel fuse network based on vision-conditioned reasoning and bilinear attentions for medical image visual question answering. *The Journal of Supercomputing* 79, 1 (2023), 1–28.

[20] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, et al. 2018. Universal Sentence Encoder. arXiv preprint arXiv:1803.11175. https://arxiv.org/abs/1803.11175

[21] Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7139–7159. https://doi.org/10.18653/v1/2022.emnlp-main.481

[22] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2667–2682. https://doi.org/10.18653/v1/2023.findings-emnlp.176

[23] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the Devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC 2014)*. BMVA Press, 6.1–6.12. https://doi.org/10.5244/C.28.6

[24] Jiuhai Chen, Jianwei Yang, Haiping Wu, Dianqi Li, Jianfeng Gao, Tianyi Zhou, and Bin Xiao. 2024. Florence-VL: Enhancing vision-language models with generative vision encoder and depth-breadth fusion. arXiv preprint arXiv:2412.04424. https://arxiv.org/abs/2412.04424

[25] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10800–10809.

[26] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9962–9971.

[27] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023. PaLI-X: Scaling up language-image pretraining. arXiv preprint arXiv:2305.18565.

[28] 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. arXiv preprint arXiv:2310.09199.

[29] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. PaLI: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794.

[30] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1290–1299.

[31] Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. 2020. Visual question answering on 360deg images. In *IEEE/CVF Winter Conference on Applications of Computer Vision*. 1607–1616.

[32] Iqbal Chowdhury, Kien Nguyen Thanh, Sridha Sridharan, et al. 2023. Video question answering for surveillance. *Authorea Preprint* (2023).

[33] HyungWon Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

[34] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 4069–4082. https://doi.org/10.18653/v1/D19-1418

[35] David Doermann and Karl Tombre. 2014. *Handbook of Document Image Processing and Recognition*. Springer Publishing Company, Incorporated.

[36] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2625–2634.

[37] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[38] Jon Mitchell. 2011. Google: Our New Search Strategy is to Compute Answers, Not Links. Retrieved March 8, 2011 from https://readwrite.com/google-our-new-search-strategy-is-to-compute-answers-not-links/

[39] Paul Freedman. 2021. *Why Food Matters*. Yale University Press.

[40] 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 (EMNLP 2016)*. Association for Computational Linguistics, 457–468. https://doi.org/10.18653/v1/D16-1044

[41] 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 answering. In *Advances in Neural Information Processing Systems* 28 (2015), 2296–2304.

[42] 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 *European Conference on Computer Vision (ECCV'18)*. 469–485.

[43] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact bilinear pooling. In *IEEE Conference on Computer Vision and Pattern Recognition*. 317–326.

[44] 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.

[45] 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 *IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[46] Bert F. Green Jr, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: An automatic question-answerer. In *Papers Presented at the May 9–11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*. 219–224.
[47] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. arXiv:2112.08614. Retrieved from https://arxiv.org/abs/2112.08614
[48] Dalu Guo, Chang Xu, and Dacheng Tao. 2021. Bilinear graph networks for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems* 34, 2 (2021), 1023–1034.
[49] Deepak Gupta, Swati Suman, and Asif Ekbal. 2021. Hierarchical deep multi-modal network for medical visual question answering. *Expert Systems with Applications* 164, 113 (2021).
[50] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P. Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 939–948.
[51] 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 *IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.
[52] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*. PMLR, 3929–3938.
[53] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. 2021. Greedy gradient ensemble for robust visual question answering. In *IEEE/CVF International Conference on Computer Vision*. 1584–1593.
[54] Monica Haurilet, Ziad Al-Halah, and Rainer Stiefelhagen. 2019. Spase-multi-label page segmentation for presentation slides. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV'19)*. IEEE, 726–734.
[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[56] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. arXiv:2003.10286. Retrieved from https://arxiv.org/abs/2003.10286
[57] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[58] Jongkwang Hong, Sungho Park, and Hyeran Byun. 2020. Selective residual learning for visual question answering. *Neurocomputing* 402 (2020), 366–374.
[59] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. 2017. Deep 360 pilot: Learning a deep agent for piloting through 360deg sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3451–3460.
[60] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *IEEE/CVF International Conference on Computer Vision*. 10294–10303.
[61] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. arXiv:2211.09699. Retrieved from https://arxiv.org/abs/2211.09699
[62] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9590–9601.
[63] Drew Hudson and Christopher D. Manning. 2019. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems* 32 (2019), 5901–5914.
[64] Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR 2018)*. https://openreview.net/forum?id=S1Euwz-Rb
[65] 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 (CVPR 2019)*. IEEE, 6700–6709. https://doi.org/10.1109/CVPR.2019.00686
[66] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1705–1715.
[67] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A. Hasan, et al.. 2018. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018))*. LNCS Lecture Notes in Computer Science, Springer, Avignon, France.
[68] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[69] 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 *IEEE Conference on Computer Vision and Pattern Recognition*. 2901–2910.

[70] Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *IEEE International Conference on Computer Vision*. 1965–1973.

[71] Danial Kamali, Elham J. Barezi, and Parisa Kordjamshidi. 2024. NeSyCoCo: A neuro-symbolic concept composer for compositional generalization. arXiv preprint arXiv:2412.15588. https://arxiv.org/abs/2412.15588

[72] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2022. Webly supervised concept expansion for general purpose vision models. In *European Conference on Computer Vision*. Springer, 662–681.

[73] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *IEEE/CVF International Conference on Computer Vision*. 1780–1790.

[74] Isaak Kavasidis, Carmelo Pino, Simone Palazzo, Francesco Rundo, Daniela Giordano, P. Messina, and Concetto Spampinato. 2019. A saliency-based convolutional neural network for table and chart detection in digitized documents. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*. Springer, 292–302.

[75] Fucai Ke, Zhixi Cai, Simindokht Jahangard, Weiqing Wang, Pari Delir Haghighi, and Hamid Rezatofighi. 2025. HYDRA: A hyper agent for dynamic compositional visual reasoning. In *European Conference on Computer Vision*. Springer, 132–149.

[76] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. 2019. Reflective decoding network for image captioning. In *IEEE/CVF International Conference on Computer Vision*. 8888–8897.

[77] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems (NeurIPS 2018)* 31 (2018), 1571–1581. https://papers.nips.cc/paper/7429-bilinear-attention-networks

[78] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*. Toulon, France. https://openreview.net/forum?id=r1rhWnZkg

[79] Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer, et al. 2020. Towards visual dialog for radiology. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. 60–69.

[80] 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.

[81] Gouthaman Kv and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 18–34.

[82] Yash Kumar Lal and Mohaddeseh Bastan. 2022. SBU figures it out: Models explain figurative language. In *3rd Workshop on Figurative Language Processing (FLP'22)*. 143–149.

[83] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data* 5, 1 (2018), 1–10.

[84] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[85] Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. 2024. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. arXiv:2405.05949. Retrieved from https://arxiv.org/abs/2405.05949

[86] Yue Li, Jin Liu, and Shengjie Shang. 2021. WMA: A multi-scale self-attention feature extraction network based on weight sharing for VQA. *Journal on Big Data* 3, 3 (2021), 111.

[87] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2928–2937.

[88] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.

[89] Shanchuan Lin, Anran Wang, and Xiao Yang. 2024. SDXL-Lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929. https://arxiv.org/abs/2402.13929

[90] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[91] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems* 35 (2022), 10560–10571.

[92] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine* 144 (2023).

[93] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI'21)*. IEEE, 1650–1654.

[94] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention correctness in neural image captioning. In *AAAI Conference on Artificial Intelligence*, Vol. 31.

[95] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*. Association for Computational Linguistics, 12789–12804. https://doi.org/10.18653/v1/2023.acl-long.714

[96] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *Advances in Neural Information Processing Systems* 36 (2024), 4377–4393.

[97] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. 2024. Prismer: A vision-language model with multi-task experts. *Transactions on Machine Learning Research (TMLR)*. https://openreview.net/forum?id=R7H43YD6

[98] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

[99] Junyu Lu, Ruyi Gan, Dixiang Zhang, Xiaojun Wu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2023. Lyrics: boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. arXiv preprint arXiv:2312.05278. https://arxiv.org/abs/2312.05278

[100] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems* 29 (2016), 289–297.

[101] Pan Lu, Liang Qiu, Jiaqi Chen, Tony 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 *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. Datasets and Benchmarks Track. https://arxiv.org/abs/2110.13214

[102] Siyu Lu, Mingzhe Liu, Lirong Yin, Zhengtong Yin, Xuan Liu, and Wenfeng Zheng. 2023. The multi-modal fusion in visual question answering: A review of attention mechanisms. *PeerJ Computer Science* 9 (2023), e1400.

[103] Zhiyang Ma, Wenfeng Zheng, Xiaobing Chen, and Lirong Yin. 2021. Joint embedding VQA model based on dynamic word vector. *PeerJ Computer Science* 7 (2021), e353.

[104] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems* 27 (2014), 1682–1690.

[105] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *IEEE International Conference on Computer Vision*. 1–9.

[106] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. https://openreview.net/forum?id=rJgMlhRctm

[107] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14111–14121.

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

[109] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *IEEE/CVF Winter Conference on Applications of Computer Vision*. 2200–2209.

[110] Aakansha Mishra, Ashish Anand, and Prithwijit Guha. 2024. Visual question answering with cascade of self- and co-attention blocks. In *Pattern Recognition, Lecture Notes in Computer Science* 14426 (2024), 20–36. https://doi.org/10.1007/978-3-031-78495-8_2

[111] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR'19)*. IEEE, 947–952.

[112] Azhan Mohammed. 2022. ResAttUNet: Detecting marine debris using an attention activated residual UNet. arXiv preprint arXiv:2210.08506. https://arxiv.org/abs/2210.08506

[113] Heejeong Nam and Jinwoo Ahn. 2024. Visual contexts clarify ambiguous expressions: A benchmark dataset. arXiv preprint arXiv:2411.14137. https://arxiv.org/abs/2411.14137

[114] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. 299–307.

[115] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. 2019. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer, 522–530.

[116] Binh X. Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. 2022. Coarse-to-fine reasoning for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4558–4566.

[117] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6087–6096.

[118] Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. 2021. Distilling knowledge in causal inference for unbiased visual question answering. In *2nd ACM International Conference on Multimedia in Asia*. 1–7.

[119] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics*. 311–318.

[120] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

[121] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

[122] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[123] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland. 2497–2511. https://doi.org/10.18653/v1/2022.findings-acl.196

[124] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *AAAI Conference on Artificial Intelligence*, Vol. 32.

[125] 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.

[126] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[127] Tanzila Rahman, Shih-Han Chou, Leonid Sigal, and Giuseppe Carenini. 2021. An improved attention for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1653–1662.

[128] Keshav S. Rawat and Sandeep K. Sood. 2021. Knowledge mapping of computer applications in education using CiteSpace. *Computer Applications in Engineering Education* 29, 5 (2021), 1324–1339.

[129] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* 28 (2015), 91–99.

[130] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*. 2662–2670. https://doi.org/10.24963/ijcai.2017/371

[131] Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. 2024. Illusory VQA: Benchmarking and enhancing multimodal models on visual Illusions. arXiv preprint arXiv:2412.08169. https://arxiv.org/abs/2412.08169

[132] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*. Springer, 146–162.

[133] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*. 618–626.

[134] Himanshu Sharma and Anand Singh Jalal. 2021. A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing* 116 (2021), 104327.

[135] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[136] Xiang Shen, Dezhi Han, Zihan Guo, Chongqing Chen, Jie Hua, and Gaofeng Luo. 2022. Local self-attention in transformer for visual question answering. *Applied Intelligence* 52, 1 (2022), 1–18.

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

[138] Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2016. The color of the cat is gray: 1 million full-sentences visual question answering (FSVQA). arXiv preprint arXiv:1609.06657. https://arxiv.org/abs/1609.06657

[139] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*. Springer, 746–760.

[140] Joseph Suarez, Justin Johnson, and Fei-Fei Li. 2018. DDRprog: A clevr differentiable dynamic reasoning programmer. arXiv preprint arXiv:1803.11361. https://arxiv.org/abs/1803.11361

[141] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 217–223.

[142] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy. 6418–6428. https://doi.org/10.18653/v1/P19-1644

[143] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China. 5100–5111. https://doi.org/10.18653/v1/D19-1514

[144] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI Conference on Artificial Intelligence*, Vol. 37. 13636–13645.

[145] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19254–19264.

[146] Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. 715–729. https://doi.org/10.18653/v1/2022.emnlp-main.45

[147] 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*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. 951–967. https://doi.org/10.18653/v1/2022.findings-emnlp.67

[148] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.

[149] Jiří Vyskočil and Lukáš Picek. 2023. VinVL+ L: enriching visual representation with location context in VQA. In *26th Computer Vision Winter Workshop*.

[150] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19175–19186.

[151] Zhonghao Wang, Kai Wang, Mo Yu, Jinjun Xiong, Wen-mei Hwu, Mark Hasegawa-Johnson, and Humphrey Shi. 2021. Interpretable visual reasoning via induced symbolic space. In *IEEE/CVF International Conference on Computer Vision*. 1878–1887.

[152] Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, et al. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025)*. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2410.12705

[153] Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. 2024. Omni-SMoLA: Boosting generalist multimodal models with soft mixture of low-rank experts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14205–14215.

[154] 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, 1 (2017), 21–40.

[155] Palmerm WuZ. 1994. VerbsSemanticsandLexicalSelection. In *32nd Annual Meeting on Association for Computational Linguistics. Las Cruces: AssociationforComputationalLinguistics*, Vol. 133. 138.

[156] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4818–4829.

[157] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal Pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201

[158] Feng Yan, Wushouer Silamu, and Yanbing Li. 2022. Deep modular bilinear attention network for visual question answering. *Sensors* 22, 3 (2022), 1045.

[159] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI Conference on Artificial Intelligence*, Vol. 36. 3081–3089.

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

[161] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems* 31 (2018), 1031–1042.

[162] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=Ee277P3AYC

[163] Zhou Yu, Xuecheng Ouyang, Zhenwei Shao, Meng Wang, and Jun Yu. 2023. Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*. IEEE, 14974–14983. https://doi.org/10.1109/CVPR52729.2023.01438

[164] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6281–6290.

[165] Desen Yuan. 2021. Language bias in visual question answering: A survey and taxonomy. arXiv preprint arXiv:2111.08531. https://arxiv.org/abs/2111.08531

[166] Dongxiang Zhang, Rui Cao, and Sai Wu. 2019. Information fusion in visual question answering: A survey. *Information Fusion* 52, 1 (2019), 268–280.

[167] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[168] Sheng Zhang, Min Chen, Jincai Chen, Fuhao Zou, Yuan-Fang Li, and Ping Lu. 2021. Multimodal feature-wise co-attention method for visual question answering. *Information Fusion* 73, 1 (2021), 1–10.

[169] Weifeng Zhang, Jing Yu, Yuxia Wang, and Wei Wang. 2021. Multimodal deep fusion for image question answering. *Knowledge-Based Systems* 212 (2021), 106639.

[170] Xinsong Zhang, Yan Zeng, Jipeng Zhang, and Hang Li. 2023. Toward building general foundation models for language, vision, and vision-language understanding tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 551–568. https://doi.org/10.18653/v1/2023.findings-emnlp.40

Received 23 November 2023; revised 8 January 2025; accepted 1 April 2025