SimpsonsVQA: Enhancing Inquiry-Based Learning with a Tailored Dataset

Anonymous WACV Algorithms Track submission

Paper ID 1100

Abstract

Visual Question Answering (VQA) has emerged as a promising area of research to develop AI-based systems for enabling interactive and immersive learning. Numerous VQA datasets have been introduced to facilitate the conventional task where individuals ask questions and the system provides answers; however, none of them support scenarios where (i) an individual asks an irrelevant question about an image, or the inverse scenario where (ii) a user answers a question that the system must assess (e.g., as correct or incorrect). Hence, in this paper, we present "SimpsonsVQA", a novel dataset for VQA derived from The Simpsons TV show, designed to promote inquiry-based learning. Our dataset is specifically designed to address not only the traditional VQA task but also the two scenarios mentioned above. It aims to cater to educational applications, harnessing the visual content of "The Simpsons" to create engaging and informative interactive systems. Furthermore, the dataset can serve as a benchmark for evaluating the performance of existing VQA models and inspire the development of novel approaches in computer vision and natural language processing. SimpsonsVQA contains approximately 23K images, 104K QA pairs, and 312K judgments (`http://simpsonsvqa.org`). We anticipate that SimpsonsVQA will inspire further research, innovation, and advancements in educational VQA.

Figure 1. Examples from the SimpsonsVQA dataset.

1. Introduction

Visual Question Answering (VQA) is a promising research field that lies at the intersection of Computer Vision (CV) and Natural Language Processing (NLP) to enable machines to answer questions about visual content [5, 18, 37, 44, 57]. The research interest in VQA has encouraged the creation of numerous datasets for constructing and evaluating VQA models including VQA v1.0 [5], VQA v2.0 [19], and GQA [23]. In addition, many datasets are purposefully crafted for specialized applications in practical domains such as healthcare [1, 21, 32], diagnosing medical images [16, 34], cultural heritage [15, 46], aiding customer service [6], enhancing entertainment experiences [17], and generating captions for social media content [51].

Despite the keen interest in VQA, most of the datasets listed above are only suitable for the scenario where an individual asks a *relevant* question about the content of an image, e.g., with the purpose of providing assistance to visually impaired persons [5, 19]. We argue in this paper that existing datasets and the current literature have overlooked two important VQA scenarios: (a) when an individual asks an *irrelevant* question about an image, and (b) the inverse scenario where an individual provides an answer to a question related to an image that the system must evaluate (e.g., "Correct", "Incorrect", or "Ambiguous"). These scenarios are particularly relevant for individuals with cognitive impairments and within educational contexts, especially early-age education. In these settings, early-age learners may not only ask coherent questions but also pose irrelevant or inconsistent ones and also provide incorrect answers related to visual content. The objective is to design engaging and informative interactive systems capable of promoting and supporting inquiry-based learning.

WACV #1100

In this paper, we present "SimpsonsVQA", a new VQA dataset that can be used to address the scenarios described above, fostering the development of intelligent systems that promote learning for individuals with cognitive disabilities and within early-age education. SimpsonsVQA is derived from The Simpsons TV show and aims at leveraging natural inclination towards cartoons, which are visually stimulating and appealing due to character identification, emotional expression, and active engagement. Each image, question, and answer triple in the dataset has been meticulously crafted using a combination of automated methods, including captioning models and ChatGPT. Then, each triple was assessed and evaluated by three workers using the Amazon Mechanical Turk (AMT) platform.

To capture the previous scenarios, SimpsonsVQA incorporates multiple triples consisting of images, questions, and answers, along with evaluative judgments. For example, in Figure 1, we show a sample of images along with free-form, open-ended, natural-language questions about the images, as well as their corresponding natural-language answers. For instance, the image in Figure 1a is linked to a relevant question and a correct answer. The image in Figure 1b has a relevant question and an ambiguous or partially correct answer. Meanwhile, the image in Figure 1c is associated with a relevant question but an incorrect answer, and the image in Figure 1d is paired with an irrelevant question. In total, SimpsonsVQA contains approximately 23K images, 104K QA pairs, and 312K judgments.

We summarize the key contributions of this work as follows: (i) *we introduce two new VQA tasks*, specifically focused on assessing answers and identifying invalid questions, which hold particular relevance in educational applications; (ii) *we present "SimpsonsVQA"*, a new VQA dataset that is explicitly tailored to address not only the traditional VQA task but also the aforementioned tasks, with the aim of fostering inquiry-based learning; (iii) *we conduct a comprehensive evaluation* with the aim to benchmark the SimpsonsVQA dataset using state-of-the-art mechine learning models. We foresee that the release of SimpsonsVQA will catalyze further research, foster innovation, and drive advancements in the field of educational VQA.

2. Related Work

Below, we provide a brief overview of existing VQA datasets and highlight some of its major applications.

Existing datasets: Several public VQA datasets have been introduced for research purposes, and Table 1 provides an exhaustive list of these datasets. They serve as invaluable resources for training and evaluating VQA models, offering a wide range of question types and varying levels of difficulty. Additionally, the images included in these datasets present a wide range of contexts.

Table 1. Popular VQA datasets. VIP - Visual Impired People; AG - Answer Grounding; OE: Open Ended; MC: Multi-Choice.

	Name	Year	Domain	#Images	#Questions	Type	Automated
1	VQA v1.0 [5]	2015	General	204,721	614,163	OE&MC	No
2	VQA v1.0 [5]	2015	Abstract sense	50,000	150,000	OE&MC	No
3	COCO-QA [44]	2015	General	123,287	117,684	OE	**Yes**
4	Binary-VQA [57]	2015	Abstract sense	50,000	150,000	MC	No
5	FM-IQA [14]	2015	General	158,392	316,193	OE	No
6	KB-VQA [53]	2015	KB-VQA	700	2,402	OE	No
7	VG [31]	2016	General	108,077	1,700,000	OE	Yes
8	SHAPE [4]	2016	Abstract shapes	15,616	244	MC	Yes
9	Art-VQA [46]	2016	Cultural Heritage	16	805	OE	No
10	FVQA [52]	2017	KB-VQA	1,906	4,608	OE	**Yes**
11	DAQUAR [38]	2017	General	1,449	12,468	OE	**Yes**
12	Visual7W [58]	2017	General	47,300	327,939	MC	Yes
13	VQA v2.0 [19]	2017	General	200,000	1,100,000	OE&MC	No
14	CLEVR [26]	2017	3D shapes	100,000	853,554	OE	Yes
15	VQA-CP1 [3]	2017	General	205,000	370,000	OE	No
16	VQA-CP2 [3]	2017	General	219,000	658,000	OE	No
17	AD-VQA [24]	2017	Advertisement	64,832	202,090	OE	No
18	VQA-MED-18 [21]	2018	Medical	2,866	6,413	OE	**Yes**
19	VQA-RAD [32]	2018	Medical	315	3,515	OE	No
20	VizWiz [20]	2018	VIP	32,842	32,842	OE	No
21	VQA-MED-19 [2]	2019	Medical	4,200	15,292	OE	Yes
22	TextVQA [47]	2019	Text-VQA	28,408	45,336	OE	No
23	OCR-VQA [41]	2019	Text-VQA	207,572	1,002,146	OE	**Yes**
24	STE-VQA [55]	2019	Text-VQA	21,047	23,887	OE	No
25	ST-VQA [9]	2019	Text-VQA	22,020	30,471	OE	No
26	OK-VQA [39]	2019	KB-VQA	14,031	14,055	OE	No
27	GQA [23]	2019	General	113,000	22,000,000	OE	**Yes**
28	LEAF-QA [10]	2019	FigureQA	240,000	2,000,000	OE	**Yes**
29	DOC-VQA [40]	2020	Text-VQA	12,767	50,000	OE	No
30	AQUA [15]	2020	Cultural Heritage	21,383	32,345	OE	**Yes**
31	RSVQA-low [20]	2020	Remote Sensor	772	77,232	OE	**Yes**
32	RSVQA-high [20]	2020	Remote Sensor	10,659	1,066,316	OE	**Yes**
33	VQA-MED-20 [1]	2020	Medical	5,000	5,000	OE	Yes
34	RadVisDial [29]	2020	Medical	91,060	455,300	OE	Yes
35	PathVQA [22]	2020	Medical	4,998	32,799	OE	Yes
36	VQA-MED-21 [7]	2021	Medical	5,500	5,500	OE	Yes
37	SLAKE [35]	2021	Medical	642	14,000	OE	No
38	GeoQA [12]	2021	Geometry Problems	5,010	5,010	MC	No
39	VisualMRC [49]	2021	Text-VQA	10,197	30,562	OE	Yes
40	A-OKVQA [45]	2022	KB-VQA	23,700	37,687	OE	No
41	VizWiz-Ground [11]	2022	VIP + AG	9,998	9,998	-	Yes
42	WSDM Cup [50]	2023	AG	45,119	45,119	-	No

The VQA v1.0 dataset [5] is often credited with popularizing the VQA task. It was introduced as one of the first large-scale VQA datasets, containing real-world images paired with open-ended questions, and it has been widely used to develop and benchmark VQA models. Later, it was expanded and improved in VQA v2.0 [19] to address language bias, establishing itself as the benchmark dataset for the VQA task. Since then, various datasets with diverse objectives have been released, such as DAQUAR [38] that focuses on indoor scenes, Visual Genome [31] and Visual7W [58] for information about the relationships between objects, CLEVR [26] in which images are rendered 3D scenes, and GQA [23] for visual reasoning and compositional answering. As shown in Table 1, some datasets have images, questions and/or answers generated and/or selected entirely through manual processes, while others involved automated procedures [21, 38, 44, 52].

VQA Applications: While VQA remains a promising and active area of research, its real-world applications are still relatively limited due to its novelty and complexity. Table 1 showcases the specific applications addressed by each VQA dataset we discussed. Initially, VQA was motivated by its potential to assist visually impaired individuals, allowing them to inquire about images and receive answers in natural language for a deeper understanding of the content [5, 19]. In addition, VQA has found promising applications in medical fields, serving as a valuable tool for analyzing medical images, supporting clinical decision-making, and diagnosing diseases [1, 2, 21, 22, 25, 29, 33, 35]. Furthermore, VQA has been explored in other domains, including Remote Sensing for extracting information from satellite images [36], cultural heritage with a focus on the old-Egyptian Amarna period [46] or painting artworks [15], and advertisement to analyze and answer questions related to advertising materials [24]. Finally, VQA has been explored for educational purposes in [12], specifically to solve geometric problems in middle school exams; however, its practical application has been hindered by its low performance compared to humans. SimpsonsVQA stands out from existing datasets by addressing not only the traditional VQA task but also introducing new scenarios, such as evaluating answers and identifying invalid questions. This unique focus caters to educational applications, enabling the development of interactive systems that encourage inquiry-based learning.

3. SimpsonsVQA Dataset

We provide in the following an overview of the SimpsonsVQA dataset. We start by detailing the methodology behind its creation, encompassing the image collection process, as well as the acquisition of questions and their corresponding answers. Following that, we will delve into the VQA tasks that are the focus of this work.

3.1. Dataset Creation

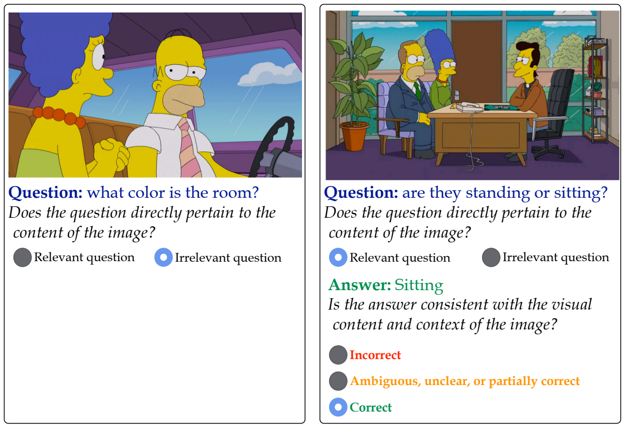
Due to the constraints imposed by limited time and budget, we adopted a pragmatic approach of automation to streamline the dataset construction process. In fact, many datasets listed in Table 1 have been created through partial automation methods [2,4,21,26,31,38,41,44,52,58]. To accomplish this, we employed a three-step approach: (1) harnessing the capabilities of Machine Learning models, particularly captioning models, to extract descriptions for each image; (2) employing ChatGPT to generate a diverse set of question-answer pairs using the obtained descriptions; and ultimately, (3) conducting a meticulous manual review by qualified workers on the AMT platform to judgments of accuracy and reliability. In the following subsections, we provide a detailed description of these steps.

Image Collection: We have collected cartoon images from the popular American sitcom, "The Simpsons". With 750 episodes spanning 34 seasons since 1989, we focused on extracting images from seasons 24 to 33. This selection includes 220 episodes, totaling approximately 80 hours of content. We used an automated process to capture images every 5 seconds, resulting in a collection of about 43,000 images. Our research team conducted a manual inspection of the images and identified approximately 1,200 inappropriate images containing violence, weapons, or sexual content, which were subsequently removed. Additionally, images lacking substantial content were also excluded. Finally, to mitigate the issue of duplicate images resulting from the fixed time interval, we employed the kNN algorithm [13] with $k = 3$ to identify and remove duplicate instances. After completing these steps, the dataset retained a total of 23,269 images.

Image Captioning: Image Captioning [51] combines CV and NLP to generate descriptive captions for images. We used the advanced pre-trained model OFA [54], known for its effectiveness in Visual Language Pre-training (VLP). OFA was trained on a dataset of 15.25 million Image-Caption samples, making it well-equipped for image captioning challenges. We fine-tuned OFA using two datasets: (a) Localized Narratives [43] and (b) Image Paragraph Captioning [30]. While the captioning model wasn't specifically trained on cartoon images, it facilitated the generation of a comprehensive, long description for each image, with an average length of approximately 300 words. Notably, it might not have captured character names or specific nuances of the TV show, yet it proved sufficient for our goal.

Generating Question-Answer Pairs: In a short period of time, ChatGPT [42] has established itself as an excellent tool for accomplishing a variety of NLP tasks. These include generating text on various subjects, acquiring information on specific topics, composing emails or messages with desired content and tone, refining text structure or wording, and more [48]. Hence, we decided to leverage the power of ChatGPT to automatically generate questions and their corresponding answers from the descriptions obtained in the previous step. In total, we prompted ChatGPT to generate a minimum of 10 question-answer pairs for each image description. After manually inspecting the generated questions and removing irrelevant ones (such as "what is the skin color of the people?" where the answer is consistently "yellow"), we obtained a comprehensive dataset comprising 103,738 image-question-answer triples.

Assessing image-question-answer triples: The accuracy and reliability of the generated questions and answers are heavily reliant on the performance of both the image captioning model and ChatGPT. However, these models are prone to errors, which can lead to the generation of irrelevant questions and/or incorrect answers. Thus, we employed the Amazon Mechanical Turk (AMT) platform to assess each image, question, and answer triple, using the interface depicted in Figure 2. In particular, we engaged human evaluators through the AMT platform and tasked them with evaluating each triple according to particular criteria. Initially, workers are presented with an image and a question, and then they are prompted to determine whether the question directly relates to the content of the image, offering a binary choice between "relevant" or "irrelevant". If the worker chooses "irrelevant", no additional action is needed for the given triple as in the case shown in Figure 2a. Otherwise, as depicted in Figure 2b, the worker must evaluate



(a) (b)

Figure 2. AMT assessment interface.

the accuracy of the answer to the question and its alignment with the image context by selecting one of these options: (i) *incorrect*, indicating that the provided answer is entirely wrong; (ii) *ambiguous or partially correct*, suggesting that the answer is unclear, open to interpretation, or it includes some correct details but also incorporates incorrect or irrelevant elements, making its validity hard to determine; and (iii) *correct*, implying that the answer is precise and directly addresses both the question and image.

To ensure the integrity of the evaluation process, each triple was assessed by three different workers. Rigorous eligibility criteria were enforced, allowing only individuals with a minimum approval rate of 99% and a track record of at least 10,000 approved HITs (Human Intelligence Tasks) to participate in the evaluation of the triples. To mitigate fraudulent or unreliable evaluations, each HIT required a minimum of 1 minute to be completed.

3.2. Task Description

Considering a set of images $\mathcal{I} = \{\mathbf{i}_1, \mathbf{i}_2, \cdots\}$, a set of questions $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots\}$, a set of possible answers $\mathcal{A} = \{a_1, a_2, \cdots\}$, the SimpsonsVQA dataset is designed to emphasize three tasks as described below.

Conventional VQA Task: Given a dataset $\mathcal{D} = \{(\mathbf{i}^{(1)}, \mathbf{q}^{(1)}, a^{(1)}), (\mathbf{i}^{(2)}, \mathbf{q}^{(2)}, a^{(2)}), \cdots, (\mathbf{i}^{(m)}, \mathbf{q}^{(m)}, a^{(m)}))\}$ with m instances, the objective of this task is to develop a classification algorithm that learns a mapping function $f : (\mathcal{I}, \mathcal{Q}) \to \mathcal{A}$, which associates each image-question pair $(\mathbf{i}^{(i)}, \mathbf{q}^{(i)})$ with its corresponding correct answer $a^{(i)}$. This initial task embodies the conventional VQA scenario, where an individual poses a question that the system is tasked to answer.

Question relevance Task: Consider a dataset $\mathcal{D} = \{(\mathbf{i}^{(1)}, \mathbf{q}^{(1)}, y^{(1)}), (\mathbf{i}^{(2)}, \mathbf{q}^{(2)}, y^{(2)}), \cdots, (\mathbf{i}^{(m)}, \mathbf{q}^{(m)}, y^{(m)}))\}$, where $y^{(i)} \in \{0, 1\}$ represents a binary label indicating the relevance of a question to an image. The objective of this task is to formulate a classification algorithm, denoted as $f : (\mathcal{I}, \mathcal{Q}) \to y$, aimed at learning a mapping that associates each image-question pair $(\mathbf{i}^{(i)}, \mathbf{q}^{(i)})$ with its corresponding binary label $y^{(i)}$. This scenario depicts an individual posing a question that is unrelated to the image.

Answer correctness Task: Consider a dataset $\mathcal{D} = \{(\mathbf{i}^{(1)}, \mathbf{q}^{(1)}, a^{(1)}, y^{(1)}), (\mathbf{i}^{(2)}, \mathbf{q}^{(2)}, a^{(2)}, y^{(2)}), \cdots, (\mathbf{i}^{(m)}, \mathbf{q}^{(m)}, a^{(m)}, y^{(m)})\}$, where $y^{(i)} \in \{\text{incorrect, ambiguous, correct}\}$ denotes a label signifying the alignment of an answer with an image-question pair. The objective is to formulate a classification algorithm denoted as $f : (\mathcal{I}, \mathcal{Q}, \mathcal{A}) \to y$, aimed at learning a mapping that associates each image-question-answer triple $(\mathbf{i}^{(i)}, \mathbf{q}^{(i)}, a^{(i)})$ with its corresponding label $y^{(i)}$. This final task represents the scenario where an individual provides an answer to a question related to an image that the system must evaluate.

Overall, we believe that the three aforementioned scenarios hold significant relevance within the realm of educational applications. Our overarching goal is to craft interactive systems that are both captivating and enlightening, fostering and facilitating inquiry-based learning experiences.

4. SimpsonsVQA Dataset Analysis

Table 2. Dataset split.

	#Image	#QA pairs
Train	13,961	77,991
Validation	3,490	12,552
Test	5,818	13,195
Total	23,269	103,738

In this section, we analyze various aspects of the SimpsonsVQA dataset, including its characteristics and distribution patterns, while also providing insights obtained from analyzing its content. As reported in Table 2, the dataset is partitioned into three subsets: train, validation, and test. It is important to note that, in order to uphold the integrity and confidentiality of the evaluation procedure, the test set remains both private and undisclosed.

4.1. Question Analysis

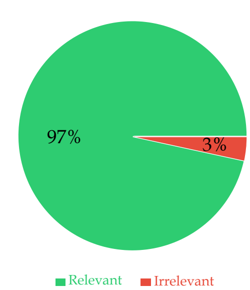


Figure 3. Ratio of question relevance as judged by AMT workers.

A total of 1,633 workers from AMT evaluated all the image-question-answer triples in our dataset. As mentioned earlier, each triple has undergone evaluation by three distinct workers, each providing judgments on two aspects: (1) the question's relevance to the image content, and (2) the accuracy of the answer in relation to the given image context. As illustrated in Figure 3, approximately 97% of the questions generated by ChatGPT (totaling 75,591 questions) have been assessed as relevant by at least 2 workers for the corresponding images. In contrast, only 3% (2,400) generated questions

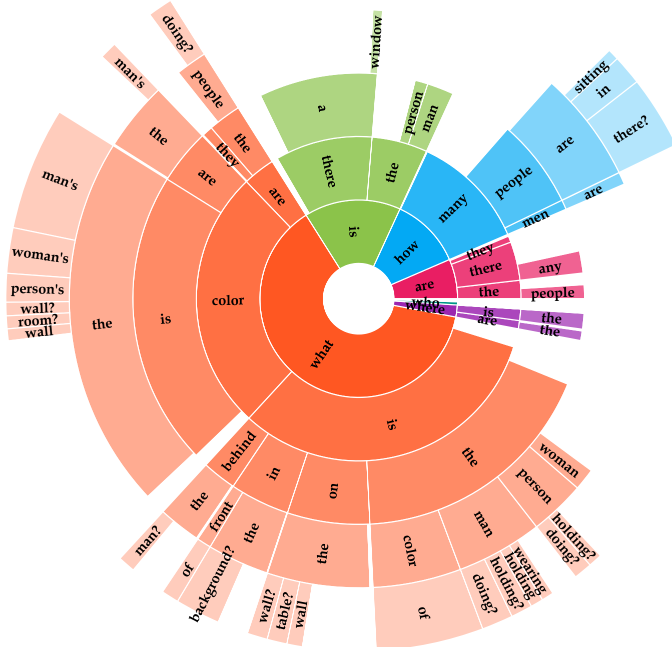


Figure 4. Distribution of first words in questions. The words are arranged in an outward radiating pattern from the center, with their ordering based on frequency. The size of the arcs corresponds to the number of questions containing each word.

lack relevance to the images.

Question Types: In Figure 4, we present an overview of the question types found in the training set. The majority of questions, approximately 62% of the questions start with the word *"what"*. Following behind are questions beginning with *"is"* and *"how"*, which account for percentages ranging from 10% to 15%. Conversely, questions initiated by words like *"are"*, *"who"*, and *"where"* make up a significantly smaller proportion. Furthermore, the most frequent question patterns include variations such as *"what is the color..."*, *"what color..."*, and *"what is the man/woman/person doing/holding"*. Additionally, a substantial number of questions involve positional inquiries, such as *"what is on/in/behind..."*, and there is also a significant presence of *"how many..."* questions.

Question Topics: As shown in Figure 5, the questions cover a wide range of topics, encompassing attribute classification 37%, object recognition 34%, counting 13%, action recognition 9%, an spatial reasoning 6%. The remaining topics collectively represent a negligible percentage, totaling only about 1% of the questions. These diverse topics play a crucial role in fostering various developmental abilities for inquiry-based learning in early-age

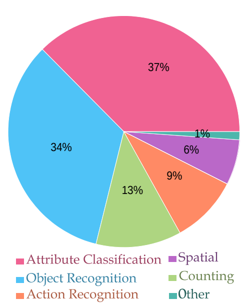


Figure 5. Question topics.

Figure 6. Word cloud of the most prominent terms in questions.

education. Figure 6 presents individual word cloud visualizations for each question type, capturing the distinctive vocabulary associated with different question categories. Each cloud highlights the frequency of specific terms, offering a visual insight into the unique linguistic characteristics of various question types.

Question Length: Figure 7 illustrates the distribution of question lengths, revealing that the majority of questions fall within the range of 3 (e.g., *"are there balloons?"*) to 10 words, with a median of 6 words. Notably, the longest question observed has a length of 18 words.

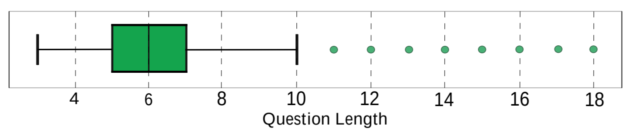


Figure 7. Question length by number of words.

4.2. Answer Analysis

Figure 8 illustrates that approximately 74% of the triples were assessed as *"Correct"* by at least two workers, while roughly 20% were deemed *"Incorrect"* by at least two workers, and around 8% were labeled as *"Ambiguous"* by at least two workers.

In Figure 9, we notice that around 45,000 triples were judged *"Correct"* by all three workers, and more than 15,000 triples were judged *"Correct"* from exactly two workers. On the other hand, around 5,500 triples were unanimously judged as *"Incorrect"* by all workers, and approximately 7,500 triples had agreement from exactly two workers labeling them as *"Incorrect"*. The number of triples judged partially ambiguous or partially correct was small.

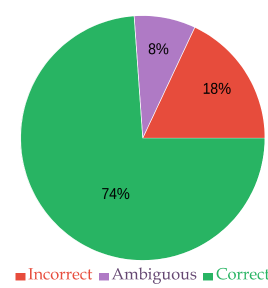


Figure 8. Worker Judgments on Triples.

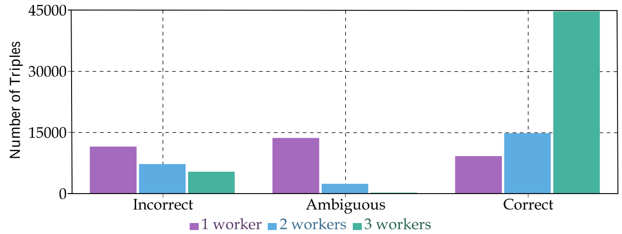
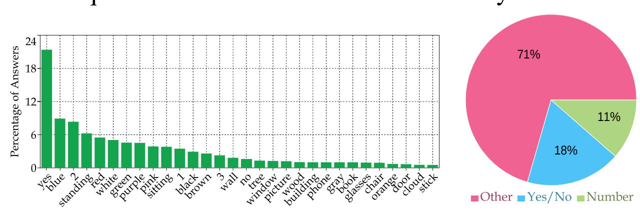


Figure 9. Assessment of triples by workers.

Popular Answers: All answers within the dataset consist of a single word. Figure 10a displays the top 30 answers with the highest frequency in the training set. Notably, the answer "yes" predictably holds the top position, constituting 20% of the answers, maintaining a notable lead of 12% over the second-place answer, "blue". Among the 15 most frequent answers, excluding "yes" and "no," the majority tend to revolve around numbers or colors. This characteristic makes the dataset particularly suitable for educational applications. Finally, as shown in Figure 10b, 18% of the questions prompted "yes" or "no" responses, while 12% of the questions received numerical answers. The remaining 70% of questions were answered in diverse ways.



(a) 30 most frequent answers. (b) Answer types.

Figure 10. Answer distribution and statistics.

Answers and Question Types: Figure 11 shows how different question types are answered. Questions starting with words like "are", "can", "do", "does", and "is" are mostly answered with either "Yes/No". Surprisingly, questions beginning with "How" not only receive numeric answers but also frequently involve words like "many" and "groups". On the other hand, questions starting with "what", "where", and "who" have a wider variety of possible answers.

5. Experimental Evaluation

In this section, we assess the effectiveness of various deep-learning models for the tasks outlined in Section 3.2, utilizing the SimpsonsVQA dataset.

5.1. Experimental setup

Baselines: To benchmark SimpsonsVQA, we have used six state-of-the-art models: (1) **LSTM Q + I [5]:** the question and image are encoded using a 2-layer LSTM and a VGGNet, respectively, and subsequently, the embeddings from both modalities are combined through element-wise multiplication; (2) **SAN [56]:** this model uses attention to focus on image regions relevant to the question. It creates a semantic representation of the question and queries the image for relevant regions, iteratively refining the answer; (3) **MLB [27]:** this model utilizes Multimodal Low-rank Bilinear Attention Networks with a low-rank bilinear pooling mechanism; (4) **MLB+Att [27]:** this variant incorporates an efficient attention mechanism within the low-rank bilinear pooling of the MLB model; (5) **MUTAN [8]:** a multimodal tensor-based Tucker decomposition that efficiently parametrizes bilinear interactions between visual and textual representations; and (6) **MUTAN+Att:** this variant extends the MUTAN model by incorporating attention.

We evaluate these six models on our three tasks, making minor adaptations to fit each task. Specifically, a notable distinction arises in the Answer Correctness task, which involves three inputs –image, question, and answer– rather than two. To achieve this, for each model, we pass the answer to a word-embedding layer followed by a dense layer. The obtained embedding is subsequently merged with the question embedding through element-wise multiplication.

Metrics: We employ the standard accuracy metric as our primary evaluation criterion. Additionally, recognizing the class imbalance for the Question Relevance task and the Answer Correctness task, we also utilize the Precision, Recall, F1-score, and AUC score to provide a comprehensive assessment of model performance.

Implementation details: All models were implemented according to original implementation. We used an image size of 480x480 pixels while fine-tuning the pre-trained models on SimpsonsVQA. All models were implemented using PyTorch 1.9 and performed on a Linux Ubuntu 18.04.1 LTS Dual Intel(R) Xeon(R) Silver CPU @2.20GHz with a GPU NVIDIA Tesla V100. All models were trained using the ADAM optimizer [28] with 50 epochs.

5.2. Results of the Conventional VQA Task

SimpsonsVQA dataset: We curated a dataset that includes only triples for which at least 2 workers have assessed as "Correct", ensuring the inclusion of only high-quality triples for evaluation. Table 3 displays the size of the resulting dataset.

Table 3. Data for the Answer Correctness task.

Dataset	#QA Pairs
Train	49,775
Validation	7,986
Test	8,561
Total	66,322

Results analysis: Table 4 presents the performance obtained on the Conventional VQA Task. In brief, we observe that Mutan+Att and MLB+Att demonstrated superior capabilities, achieving accuracies of 72.86% and 72.00%, respectively. Meanwhile, the MLB and Mutan models achieved lower overall accuracies of 57.86% and 58.28%, indicating that the inclusion of attention mechanisms significantly improved the models' overall performance.

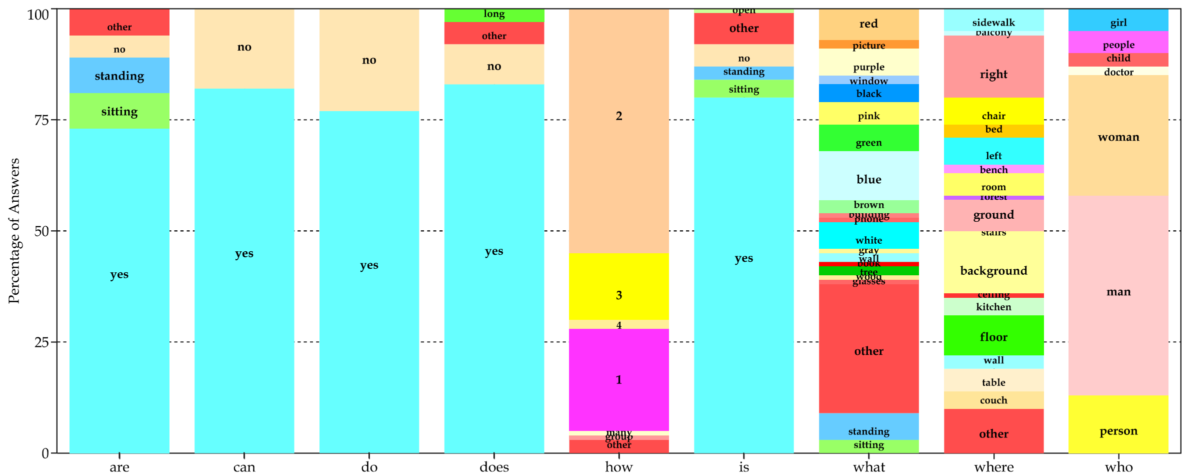


Figure 11. Answer distribution based on first words of questions.

Table 4. Performance on the Conventional VQA Task.

Model	Accuracy			
	Number	Yes/No	Other	All
LSTM Q + I	0.6980	0.9497	0.5217	0.6441
SAN	0.7240	0.9540	0.5591	0.6697
MLB	0.6442	0.9541	0.4456	0.5786
MLB+Att	**0.7311**	0.9545	0.6446	0.7200
Mutan	0.6276	0.9462	0.4472	0.5828
Mutan+Att	0.7227	**0.9562**	**0.6539**	**0.7286**

The models displayed remarkable proficiency in answering "Yes/No" questions, with all models achieving a strong accuracy range of 94%-96%. These questions are evidently easier for the models due to their straightforward binary nature. On the other hand, "Number" questions are relatively manageable for the models, with MLB-Att and SAN achieving accuracies of 73.11% and 72.40%. However, all models encountered challenges in accurately answering "Other" questions, which involve open-ended responses and require deeper understanding.

5.3. Results of the Question Relevance Task

SimpsonsVQA dataset: We curated a dataset that comprises images and questions labeled as relevant or irrelevant, determined through the majority decision of the workers. As indicated in Table 5, the dataset is imbalance with a relevant-to-irrelevant ratio of 97:3.

Table 5. Data for the Question Relevant task.

Dataset	#Relevant QA Pairs	#Irrelevant QA Pairs	Total
Train	75,591	2,400	77,991
Validation	12,146	406	12,552
Tests	12,813	382	13,195
Total	100,550	3,188	103,738

Results analysis: Table 6 gives an overview of baseline models' performance. None of the models exceeded the accuracy of the majority vote classifier. All models achieved accuracy ranging from 87.70% to 88.39%, with Mutan+Att achieving the highest accuracy at 88.39%. Also, we observe that Mutan+Att and MLB+Att stand out, capturing more relevant and irrelevant instances. The F1-Scores for these

Table 6. Performance on the Question Relevance Task.

Model	Accuracy	AUC	Precision		Recall		F1-Score	
			Rel	Irrel	Rel	Irrel	Rel	Irrel
LSTM Q + I	0.8770	0.7359	0.8943	0.4094	0.9763	0.1242	0.9335	0.1906
SAN	0.8811	0.7235	0.8945	0.4263	0.9776	0.1257	0.9342	0.1939
MLB	0.8779	0.7259	0.8955	**0.4268**	0.9757	0.1361	0.9339	0.2050
MLB+Att	0.8839	0.7402	0.9049	0.3723	0.9454	0.2459	0.9247	**0.2962**
Mutan	0.8782	0.7245	0.8941	0.4224	0.9779	0.1217	0.9341	0.1887
Mutan+Att	0.8725	**0.7414**	0.9020	0.4057	0.9599	**0.2087**	0.9301	0.2758
Majority Vote	**0.9710**	0.5	**0.9710**	0	**1**	0	**0.9853**	0

models further corroborate their better balanced performance. Finally, Mutan+Att achieves the highest AUC score of approximately 0.7414, followed closely by MLB+Att with a score of around 0.7402. These two models exhibit strong discriminatory power in their predictions.

5.4. Results of the Answer Correctness Task

SimpsonsVQA dataset: We constructed the dataset following these guidelines: When there was unanimous consensus among two or more workers, the majority perspective was assigned as the label for the triple. If unanimous agreement is not reached, we assign the label as "Ambiguous". Table 7 provides dataset details, where C stands for "Correct", AM for "Ambiguous", and IC for "Incorrect".

Table 7. Data for the Answer Correctness task.

Dataset	#C QA Pairs	#AM QA Pairs	#IC QA Pairs	Total
Train	59,472	5,912	12,607	77,991
Validation	9,477	960	2,115	12,552
Test	10,020	925	2,250	13,195
Total	78,969	7,797	16,972	103,738

Results analysis: Table 8 provides a summary of the performance of the baseline models. Given that the data predominantly consists of "Correct" triples, the models achieve high performance for this category. In contrast, the performance for the "Ambiguous" category is often notably low. The performance of the "Incorrect" class experiences substantial fluctuations, ranging from 17.02% to 40.26%. MLB+Att and Mutant+Att stand out as the models showcasing the highest proficiency in classifying Incorrect Triples. The obtained results reaffirm MLB+Att and

Table 8. Performance on the Answer Correctness Task.

Model	Accuracy	Precision			Recall			F1-Score		
		C	AM	IC	C	AM	IC	C	AM	IC
LSTM Q + I	0.7660	0.7835	0.000	0.5107	0.9660	0.0000	0.1907	0.8652	0.0000	0.2777
SAN	0.7712	0.8023	0.025	0.5172	0.9476	0.0001	0.3067	0.8689	0.0018	0.3850
MLB	0.7672	0.7750	0.0000	**0.5604**	**0.9735**	0.0000	0.1190	0.8669	0.0000	0.1963
MLB+Att	0.7530	0.8119	0.1753	0.4736	0.9078	0.0663	0.3465	0.8571	0.0960	0.4001
Mutan	**0.7742**	0.7969	0.1570	0.5459	0.9586	0.0051	0.2691	**0.8703**	0.0009	0.3904
Mutan+Att	0.7500	**0.8180**	**0.1909**	0.4601	0.8968	**0.0819**	**0.3710**	0.8556	**0.1139**	**0.4107**

Mutan+Att as robust models, likely due to their attention mechanism and suitable architectures.

5.5. Discussion

Drawing from our experimental findings, several key observations come to light. First, among the three VQA tasks mentioned earlier, the assessed baseline models demonstrate effectiveness in the Conventional VQA Task, while facing challenges in predicting questions that don't fall into the categories of "yes/no" or "number" questions. This outcome is comprehensible given that these models are crafted with a distinct emphasis on the conventional VQA task.

Second, in the Question Relevance task, existing models showed comparable performance. What emerges as a particularly interesting avenue for future research is the challenge of ascertaining the validity of a question in relation to the visual content depicted within the image. This aspect introduces a new layer of complexity to the problem. Essentially, it implies that the models not only need to understand the question itself, but also possess the capability to assess the appropriateness of the question based on what is observable within the given image. This novel challenge opens up opportunities for exploring innovative approaches to imbuing AI models with a deeper understanding of context and visual cues, paving the way for more sophisticated and contextually aware question-answering systems

Finally, the Answer Correctness task emerges as the most formidable challenge. Unlike the conventional VQA task, where models are required to grasp image and question content to formulate answers, this task adds an extra layer of complexity. Models must not only understand the image and question to craft responses, but they must also fathom the interplay between image, question, and answer to classify responses as correct, incorrect, or ambiguous. However, the crux of the matter lies in the models' inability to effectively classify incorrect or perplexing answers. The VQA Answer Correctness task can be seen as an elevated iteration of the standard VQA exercise, demanding a deeper and more nuanced level of understanding.

6. Potential Negative Societal Impact

While the SimpsonsVQA holds valuable educational and learning applications, there are several potential negative implications that need to be acknowledged.

Stereotyping and Bias: Since the dataset is derived from The Simpsons TV show, it may inadvertently contain stereotypes, biases, or cultural references that could perpetuate negative perceptions or reinforce existing biases. Hence, if the dataset is used in educational settings, there's a risk that learners might absorb stereotypes, incorrect information, or biased perspectives from the dataset's content.

Cognitive and Emotional Impact: The use of AI-generated content in educational contexts could impact learners' cognitive and emotional development. Ensuring that the content is age-appropriate, respectful, and conducive to positive learning experiences is paramount.

Over-reliance on AI-driven tools: There's a concern regarding the possibility of excessive dependence on AI-driven educational tools, potentially diminishing the role of educators and human interaction in the learning process. While AI can provide valuable support, it's crucial to maintain a balance that combines technological advancements with human guidance to ensure effective experiences.

7. Conclusion & Future Work

In the realm of Visual Question Answering (VQA), which bridges Computer Vision and Natural Language Processing, a surge of interest has led to the creation and evaluation of large range of datasets for diverse applications in healthcare, entertainment, customer service, and more. However, prevailing datasets often neglect scenarios where questions are irrelevant or answers need evaluation. This paper addresses these gaps by introducing the "SimpsonsVQA" dataset, tailored for educational contexts, where learners' questions and answers might vary widely in relevance and accuracy. Leveraging cartoons from The Simpsons, this dataset fosters the development of intelligent systems for early-age education. By presenting novel VQA tasks and a carefully constructed dataset, this work aims to advance inquiry-based learning and spur further research and innovation in educational VQA.

Future research includes exploring advanced techniques for automated question relevance assessment and answer validation to enhance the robustness of the system. Additionally, investigating the integration of real-time feedback mechanisms and adaptive learning strategies within the educational context could further optimize the interactive learning experience facilitated by the SimpsonsVQA dataset.

Ethical considerations: In creating SimpsonsVQA, there was no collection or publication of any personal or critical data related to the AMT workers. The annotators responsible for labeling the SimpsonsVQA dataset were compensated fairly for their efforts, adhering to the minimum wage standards set by the platform.

[1] https://creativecommons.org/licenses/by-nc-sa/4.0/

References

[1] Asma Ben Abacha, Vivek V Datla, Sadid A Hasan, Dina Demner-Fushman, and Henning Müller. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *CLEF (Working Notes)*, 2020. 1, 2, 3

[2] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *CLEF (working notes)*, 2(6), 2019. 2, 3

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 2

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016. 2, 3

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 6

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 1

[7] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania, September 21-24 2021. CEUR-WS.org. 2

[8] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 6

[9] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 2

[10] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521, 2020. 2

[11] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022. 2

[12] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021. 2, 3

[13] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. 3

[14] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015. 2

[15] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 92–108. Springer, 2020. 1, 2, 3

[16] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. Cross-modal self-attention with multi-task pretraining for medical visual question answering. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 456–460, New York, NY, USA, 2021. Association for Computing Machinery. 1

[17] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 2

[20] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 2

[21] Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *CLEF (Working Notes)*, 2018. 1, 2, 3

[22] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 2, 3

[23] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 1, 2

[24] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video

advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. 2, 3

[25] Bogdan Ionescu, Henning Müller, Renaud Péteri, Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, Serge Kozlovski, Vitali Liauchuk, Yashin Dicente Cid, et al. Overview of the imageclef 2021: Multimedia retrieval in medical, nature, internet and social media applications. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 345–370. Springer, 2021. 3

[26] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2, 3

[27] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 6

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer Beymer, et al. Towards visual dialog for radiology. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 60–69, 2020. 2, 3

[30] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*, 2017. 3

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2, 3

[32] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5, 2018. 1, 2

[33] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 3

[34] Zhihong Lin, Donghao Zhang, Qingyi Tac, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *arXiv preprint arXiv:2111.10056*, 2021. 1

[35] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 2, 3

[36] Sylvain Lobry, Jesse Murray, Diego Marcos, and Devis Tuia. Visual question answering from remote sensing images. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4951–4954. IEEE, 2019. 3

[37] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 1

[38] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014. 2, 3

[39] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2

[40] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2

[41] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952, 2019. 2, 3

[42] OpenAI. Chatgpt. OpenAI API, 2021. Accessed: [13/17/2023]. 3

[43] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 3

[44] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015. 1, 2, 3

[45] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer, 2022. 2

[46] Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the COLING 2016 Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17. ACL, 2016. 1, 2, 3

[47] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2

[48] Shahab Saquib Sohail, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor. Decoding chatgpt: A taxonomy of existing research, current challenges, and possible future directions. *Journal of King Saud University - Computer and Information Sciences*, 35(8):101675, 2023. 3

[49] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 2

[50] Dmitry Ustalov, Nikita Pavlichenko, Daniil Likhobaba, and Alisa Smirnova. WSDM Cup 2023 Challenge on Visual Question Answering. In *Proceedings of the 4th Crowd Science Workshop on Collaboration of Humans and Learning Algorithms for Data Labeling*, pages 1–7, Singapore, 2023. 2

[51] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 3

[52] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 2, 3

[53] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015. 2

[54] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3

[55] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. 2

[56] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[57] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2

[58] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 2, 3