Encyclopedic VQA:
Visual questions about detailed properties of fine-grained categories
Supplementary Material

Thomas Mensink[†,*]
mensink@google.com

Jasper Uijlings[†,*]
jrru@google.com

Lluis Castrejon[*]
lluisc@google.com

Arushi Goel[‡]
goel.arushi@gmail.com

Felipe Cadar[‡]
cadar@dcc.ufmg.br

Howard Zhou[*]
howardzhou@google.com

Fei Sha[*]
fsha@google.com

André Araujo[*]
andrearaujo@google.com

Vittorio Ferrari[*]
vittoferrari@google.com

Abstract

In this supplementary material, we provide additional details to our paper Encyclopedic-VQA. We evaluate more thoroughly the correctness of the dataset (Sec. 1), the BEM evaluation metric (Sec. 2), CLIP [12] for retrieval (Sec. 3), we provide some more qualitative results (Sec. 4), more extensive dataset statistics (Sec. 5), and finally we provide the specific prompts that were used for large language (PaLM) and vision+language (PaLI) models (Sec. 6).

1. Correctness of our dataset

We rely on iNat21 and GLDv2 to have clearly identifiable categories on the test images of our dataset, and to obtain their ground-truth labels. Furthermore, we only use categories which unambiguously map to Wikipedia articles by explicitly seeking for one-to-one mappings (see Sec.4.1 in the main paper). The main remaining possible source of error is whether a question can be answered given the corresponding Wikipedia article, and whether the recorded ground-truth answer is indeed correct. Human annotators systematically validate these aspects for every single question+answer pair (see Sec.4.2 and Sec.4.3). Yet, we perform here an additional user study to get an numerical estimate of quality in this sense.

In this study, we randomly sample 100 questions from the test set. Then we ask six experts to each answer 50 questions each, given the corresponding Wikipedia page as reference. These expert were not involved in the original data collection process. This process results in three answers per question. If the majority of the 3 expert answers matches our collected ground-truth according to BEM [5], we consider that question to be *answerable* given the Wikipedia page, and our ground-truth to be correct. We find that this holds for most of our questions (86%).

To put this number in context, we also estimate the answerability of A-OKVQA [14] with respect to their evaluation metric. A-OKVQA follows previous VQA datasets and provides 10 ground-truth answers per question. A predicted answer is counted as correct if it matches with 3 ground-truth answers (exact string matching [4, 14], after normalization such as removing punctuation and articles, converting numbers to digits, etc.). So we can consider a question is answerable if there are at least 3 equivalent answers out of the 10. We find that 86% of the A-OKVQA questions are answerable.

To conclude, this user study demonstrates that our Encyclopedic-VQA dataset is of very high quality.

2. Quality of BEM [5] evaluation measure

We want to understand how well the BERT Matching (BEM) [5] evaluation measure mirrors human judgments. To do so we build on the user study above and ask an expert human to judge whether each answer from the user study matches the collected ground-truth answer or not. The judge has access to both the question Q and the corresponding Wikipedia page.

[†] Equal contribution. [‡] Work done during internship at Google. [*] Google Research.

We find that the BEM judgements equal human judgements in the vast majority of the cases (96%). BEM is a stricter judge in 3% of the cases, the human in 1%. This demonstrates that BEM correlates very highly with human judgement of correctness of an answer, confirming what reported in [5] for other datasets.

Exact Match. Many VQA datasets perform exact string matching for their evaluation [4, 8, 11, 14]. Because it is a strict evaluation measure and because their questions are more open-ended (they are not supported by a controlled knowledge base, Sec. 4 of the main paper), those datasets collect 10 ground-truth answers per question, to cover some of the variability in answer formulation. This then enables a more relaxed evaluation, as the model answer need only match some of the 10 ground-truth ones (three matches for a perfect answer). Nevertheless, it is instructive to verify whether exact matching would work on our dataset, where we have a single ground-truth answer. We start from the publicly available exact matching implementation of [9] and include additional relaxations for number comparisons (e.g. numbers over 100 - usually years - may be off by one, ranges are correct if they partially overlap, etc). We find that exact match judgements are equivalent to human judgements only in 68% of the cases. Where they differ, exact match is always overly strict, rejecting answers that a human would judge as correct. This demonstrates that exact matching does not work well for our dataset, and justifies our choice of BEM as the evaluation measure to check whether predicted model answers match ground-truth ones in Sec. 5 of the main paper.

3. CLIP retrieval in existing retrieval-augmented VQA systems [9, 10]

KAT [9] and REVIVE [10] are two retrieval-augmented VQA systems which use frozen CLIP [12] embeddings to perform retrieval. More specifically, they first encode their text-only knowledge base (extracted from Wikidata) using the language tower of CLIP. At test time, given a VQA question, they encode its image I with CLIP and then compare its embedding to the knowledge base embeddings to retrieve the most similar entries. The top few most similar entries are then passed on to a T5 model [13] which is trained to produce the final answer. Note how correct retrieval is crucial to perform well on *Encyclopedic*-VQA, as the retrived knowledge base entries should contain the answer for the overall system to succeed. Hence, we now test the CLIP-based retrieval component of KAT [9] and REVIVE [10] in isolation, to estimate whether they would be able to succeed on *Encyclopedic*-VQA.

We represent Wikipedia articles as the CLIP embedding of their title and description strings concatenated. At test time, given a VQA question, we use CLIP to embed the image I to form the query for retrieval, proceeding as in [9, 10]. Results in Tab. 1 demonstrate that recall is low: the correct Wikipedia article corresponding to the subject of the question is retrieved in the first position only 3.3% of the time. Even within the top-20 results, retrieval accuracy is still only at 16.5%. This suggest that KAT/REVIVE would not work well on our dataset.

Lens retrieval. In Sec. 5.3 (of the main paper) we use Google Lens to identify the subject C of a VQA question in our datatset, i.e. a iNat21 or GLDv2 category, which works well (Tab. 1). Lens is an image-based system which indexes billions of images on the web and retrieves relevant ones based on their visual similarity. Hence, it has a greater chance to find a web image resembling a VQA test image, than when restricting the search only to Wikipedia. Attached to the image is often various meta-data which enables to determine the name of the subject category (which we then use to find the right Wikipedia page). Recognizing landmarks and species from the natural world are typical use-cases so we can expect Lens to work well for recognizing the subjects C in our dataset. However, academic results on iNat21 [3] and GLDv2 [2] suggests that specialized systems work very well on these datasets, suggesting that a good image-based classifier would make a viable substitute for Lens.

Generally, the core points of our paper are that (1) *Encyclopedic*-VQA poses a hard challenge for large LLMs and VLMs, and (2) solving it requires augmenting the LLM/VLM with a retrieval component to access a knowledge base. The exact choice of retrieval component is flexible and likely subject to further exploration.

	Recall			
Method	@1	@5	@10	@20
CLIP [12]	3.3%	7.7%	12.1%	16.5%
Lens [1]	47.4%	62.5%	64.7%	65.2%

Table 1: **Recall results for CLIP and Lens.** We report the recall in retrieving the right KB article within the top-K documents.

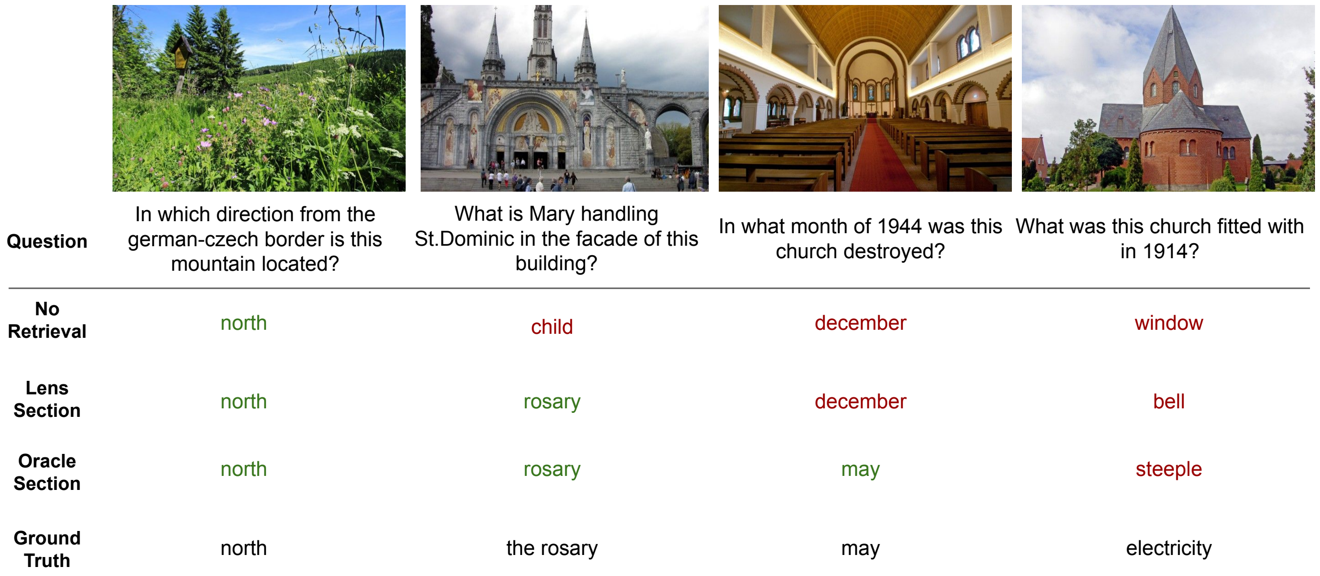
Question
In which direction from the german-czech border is this mountain located?
What is Mary handling St.Dominic in the facade of this building?
In what month of 1944 was this church destroyed?
What was this church fitted with in 1914?
No Retrieval
north
child
december
window
Lens Section
north
rosary
december
bell
Oracle Section
north
rosary
may
steeple
Ground Truth
north
the rosary
may
electricity

Figure 1: **PaLI qualitative results.** We present 4 multi-modal questions and the answers produced with different experimental setups, using the **PaLI** model: without retrieval ("No Retrieval"), Lens-based retrieval of KB section ("Lens Section"), Oracle retrieval of KB section ("Oracle Section"). Additionally, we provide the ground-truth answer in the last row. We show cases where each of the setups can produce the correct answer, as well as an example where all methods fail (the right-most one).

		Number of Q+A pairs			Number of Categories			% Unique C		Total (I, Q, A) triplets		
		Train	**Val**	**Test**	**Train**	**Val**	**Test**	**Val**	**Test**	**Train**	**Val**	**Test**
iNat21	Templated	8,133	200	500	1,983	114	300	26%	21%	40,665	1,000	500
	Automatic	96,317	1,000	1,500	4,061	343	562	5%	6%	481,585	5,000	1,500
	Multi Answer	11,262	200	500	3,585	118	318	31%	22%	56,310	1,000	500
	Two Hop	10,819	200	500	2,499	131	359	11%	6%	54,095	1,000	500
GLDv2	Templated	5,795	200	500	1,808	135	268	27%	28%	25,870	827	500
	Automatic	57,124	750	1,250	1,965	138	208	9%	9%	255,529	3,025	1,250
	Multi Answer	12,667	200	500	6,449	151	339	45%	56%	56,426	844	500
	Two Hop	10,221	200	500	1,405	99	154	11%	11%	45,771	895	500
	Total	**212,338**	**2,950**	**5,750**	**16,249**	**1,071**	**2,152**	**14%**	**17%**	**1,016,251**	**13,591**	**5,750**

Table 2: **Dataset Statistics.** We report the number of question, answer, categories and triplets for both supporting datasets and different question types. In total our dataset contains 1M VQA triplets, making it the largest of its kind.

4. Qualitative results

We present qualitative results for some of the methods we experimented with in Figures 1 and 2. These illustrate cases where the different setups may work, and cases where all of them fail. For example, given a textual question *What does this reptile eat?*, PaLM can reasonably guess an answer without any additional context: *insects* – since that's what many reptiles eat.

For more specific information, relating to more detailed properties (*e.g.*, number of eggs a specific reptile lays, dates, how big a specific fish may become), large models may still make reasonable guesses, but generally this leads to incorrect answers. Augmenting large models with context from retrieved knowledge improves the accuracy in these cases substantially, leading to precise answers that are attributable to the piece of knowledge that was retrieved.

Finally, we also see cases where the generated answer is incorrect, even if the correct piece of knowledge is retrieved. This indicates that in some cases large models may still have difficulties using retrieved knowledge to generate accurate answers.

Question	What does this reptile eat?	How many eggs does this reptile typically lay?	What does this animal eat?	How big does this fish typically become?
No Retrieval	insects	10-15	insects	6 cm
Lens Section	insects	3-6	plankton	13 cm
Oracle Section	insects	3-6	mussels	13 cm
Ground Truth	insects	3-6	mussels	25 cm

Figure 2: **PaLM qualitative results.** We present 4 multi-modal questions and the answers produced with different experimental setups, using the **PaLM** model: without retrieval ("No Retrieval"), Lens-based retrieval of KB section ("Lens Section"), Oracle retrieval of KB section ("Oracle Section"). Additionally, we provide the ground-truth answer in the last row. We show cases where each of the setups can produce the correct answer, as well as an example where all methods fail (the right-most one).

5. Dataset Statistics

In Table 2 we provide a more detailed overview of the dataset statistics. For more details see Section 4 of the main paper.

6. Prompts for dataset creation

As discussed in Section 4 of the paper, we used a FLAN [7] version of the PaLM [6] model to (1) rephrase automatically generated single-hop questions (both single-answer and multi-answer), and to (2) combine two single-hop questions into a two-hop question. Generally, we found that the model tends to work better when fed with very specific (and sometimes redundant) instructions, combined with chain-of-thought prompts [15] where possible. Additionally, we added difficult cases as examples in the prompt, to guide the model's behavior in more challenging situations.

6.1. Rephrasing automatically generated single-hop questions

We replaced the name of the category C in the question by its super category. This rephrasing process was run for all automatically generated single-hop questions (single and multi-answer). For example, the question *Where is the Église Saint-Cannat located?* was rephrased to *Where is this church located?* After several iterations, we decided to add three examples in the prompt, using chain-of-thought reasoning with three steps. The model generally writes the three required steps in the output, and we parse the rephrased question right after the third step written by the model. Sometimes, the model would return an incorrectly rephrased question. We observe that the main failure case is when the model simply copies the input question to the output. Thus, we filter out any rephrased question that is identical to the original question fed in its input. We sampled 100 rephrased questions to manually assess the quality of the rephrasings, which we found to be correct in 90% of the cases. The final prompt is given in Prompt 1, where "\$C\$" and "\$Q\$" are placeholders for the category C and the textual (pre-rephrasing) question Q.

6.2. Chaining single-hop questions into two-hop questions

We created two-hop questions by chaining two single-hop questions, where the answer to the first single-hop question serves as a bridge entity that makes a connection to the second single-hop question. For example, given the first single-hop question (SQ1) *What is the main competitor for food for this animal?* with answer (SA1) *Spotted hyena*, and the second single-hop question (SQ2) *What is the population size of this animal?* with answer (SA2) *Between 27,000 and 47,000 individuals*, we generated *What is the population size of the main competitor for food of this animal?* In this example, the entity *spotted hyena* serves as a bridge between the two single-hop questions. Note that the answer to the two-hop question is identical to the answer to the second single-hop question, *Between 27,000 and 47,000 individuals*.

After several iterations, we crafted a prompt with 4 examples of varying difficulty (Prompt 2). The output of the model is taken as the chained two-hop question. Sometimes, though, the model would return an incorrect two-hop question. For this reason, we designed a second prompt to validate the two-hop question provided as the initial output (Prompt 3). The model is asked to answer the two-hop question using the two initial single-hop questions with answers as context. If the predicted answer is identical to the answer of the second single-hop question, then the two-hop question is validated and kept in our dataset; otherwise, it is discarded. Finally, we filter some common failure cases, such as when the model outputs the exact first single-hop or second single-hop question as the chained two-hop question. We sampled 100 chained two-hop questions to manually assess their quality, which we found to be correct in 88% of the cases. In the final prompts (Prompts 2-3) “$SQ1$” / “$SA1$” are placeholders for the first single-hop question / answer, “$SQ2$” / “$SA2$” are placeholders for the second single-hop question / answer, and “Q” is the placeholder for the generated two-hop question.

7. Prompts for evaluation

In Section 5 we analyze the performance of large models in our dataset. Similarly to the dataset creation process, we use diverse prompts to adapt and improve the behavior of these models. In particular, we use prompts to incorporate retrieval results and to identify relevant sections for articles retrieved with Lens.

7.1. Prompts for question answering

We use textual Prompts 4-7 to produce answers with large models. Note that we use the same prompts when using PaLI or PaLM. However, we always use the question image $\mathcal{I}$ as an additional input to the model when using PaLI. Note that we use Q to refer to the textual part of a question, C to denote the question subject $\mathcal{C}$ name, Art to denote the full text of a Wikipedia article and S to denote a section of a Wikipedia article. We use the same retrieval prompts for Lens and Oracle setups. However, for Lens experiments, we only include Art or S if available, as we might not retrieve entities with Lens or find relevant sections with PaLM. When not available, we revert to using Prompt 4.

7.2. Prompt for relevant section identification

We use Prompt 8 to identify relevant sections in a Wikipedia article retrieved by Lens. To do so, we query PaLM for each section S in the Lens retrieved Wikipedia article for a question Q. Note that we use a few examples in the prompt to condition PaLM to generate a yes/no answer. The answer produced by PaLM is converted into a string and matched to either `yes` or `no`. When no match is found, we assume that section is not relevant to the question. If more than one section is identified as relevant for a question, the input S to Prompt 7 becomes the concatenation of all relevant sections.

Prompt 1: Rephrasing automatically-generated single hop question

```
In this task, please rephrase the question by replacing the entity name by the word "this",
followed by the type of the entity.  See the examples below:

EXAMPLE 1:
entity name:  eiffel tower
question:  How tall is the eiffel tower?
step 1 (find type of entity):  The eiffel tower is a type of:  tower
step 2 (write the word "this" followed by the type obtained in "step 1"):  this tower
step 3 (final rephrased question):  How tall is this tower?

EXAMPLE 2:
entity name:  salmon
question:  Which country is the largest producer of salmon?
step 1 (find type of entity):  The salmon is a type of:  fish
step 2 (write the word "this" followed by the type obtained in "step 1"):  this fish
step 3 (final rephrased question):  Which country is the largest producer of this fish?

EXAMPLE 3:
entity name:  Grand Beach Provincial Park
question:  grand beach provincial park is located on the east side of what lake?
step 1 (find type of entity):  The Grand Beach Provincial Park is a type of:  park
step 2 (write the word "this" followed by the type obtained in "step 1"):  this park
step 3 (final rephrased question):  this park is located on the east side of what lake?

Note that the entity name should not be part of the rephrased question.
Please make sure to write out all 3 steps as in the examples above.

Based on the above examples, provide a rephrased question for the following case:
entity name:  $C$
question:  $Q$
```

Prompt 2: Chaining two single-hop questions into a two-hop question

```
EXAMPLE 1:
question 1:  in which city is this building located?
answer 1:  San Francisco
question 2:  what is the average temperature in San Francisco?
answer 2:  15 Celsius
combined question:  What is the average temperature in the city where this building is located?

EXAMPLE 2:
question 1:  what is the predator of this animal?
answer 1:  Lion
question 2:  What is the weight of a lion on average?
answer 2:  190 kilograms
combined question:  What is the average weight of the predator of this animal?

EXAMPLE 3:
question 1:  In what country is this plant found?
answer 1:  Australia
question 2:  What does australia’s size give it a wide variety of?
answer 2:  landscapes and climates
combined question:  What does the size of the country where this plant is found give it a wide
variety of?

EXAMPLE 4:
question 1:  What country is this plant the national flower of?
answer 1:  South Africa
question 2:  south africa is a member of the commonwealth of nations and what other
organization?
answer 2:  the G20
combined question:  The country that this plant is the national flower of is a member of the
commonwealth of nations and what other organization?

Based on the above 4 examples, provide a combined question for the following case, such that
the answer to the combined question is the same as the answer to question 2:
question 1:  $SQ1$
answer 1:  $SA1$
question 2:  $SQ2$
answer 2:  $SA2$
combined question:
```

Prompt 3: Validating two-hop question given the original single-hop questions

```
question 1:  $SQ1$
answer 1:  $SA1$
question 2:  $SQ2$
answer 2:  $SA2$
Based on the questions and answers above, please answer the following question:  $Q$
```

Prompt 4: Question only evaluation

```
Question:  $Q$
The answer is:
```

Prompt 5: Retrieval with entity name

```
Entity name:  $C$
Question:  $Q$
The answer is:
```

Prompt 6: Retrieval with KB Article

```
Context:  $Art$
Question:  $Q$
The answer is:
```

Prompt 7: Retrieval with KB Section

```
Context:  $S$
Question:  $Q$
The answer is:
```

Prompt 8: Lens retrieval to get KB Section

```
Can the answer to the question be found in the text?
Question:  Is this fungus edible?
Text:  This compound induces mammalian cells (specifically, the cell line HL60 to differentiate
into granulocyte- or macrophage-like cells.  The fungus also contains the mycotoxin muscarine,
and the antifungal metabolite strobilurin D. Despite the presence of these toxins, some guides
list this fungus safe for human consumption.
The answer is:  yes

Can the answer to the question be found in the text?
Question:  In which season does this plant give flowers?
Text:  This cactus has stems about 1/2-1 inch wide with 6-9 edges.  Its flowers are white, up
to 30 centimetres in diameter with a scent redolent of vanilla.  The flowers open after sundown,
closing and wasting after a few hours.  By 9 am the next day they are gone.
The answer is:  no

Can the answer to the question be found in the text?
Question:  What is the habitat of this animal?
Text:  X is native to Europe and North Africa through to Central Asia.  It is introduced to the
United States and parts of South America.  It widespread across the northeastern United States
and eastern Canada, and can be found outside, or more commonly inside houses.  It is thought to
have been introduced into America from Europe by English colonists.
The answer is:  yes

Can the answer to the question be found in the text?
Question:  $Q$
Text:  $S$
The answer is:
```

References

[1] Google Lens. https://lens.google.com - Web interface available at https://images.google.com. 2

[2] Google universal image embedding challenge leaderboard. https://www.kaggle.com/competitions/google-universal-image-embedding/leaderboard. 2

[3] inat challenge 2021 - fgvc8 - leaderboard. https://www.kaggle.com/competitions/inaturalist-2021/leaderboard. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015. 1, 2

[5] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022. 1, 2

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways. *arXiv*, 2022. 4

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 4

[8] 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 *CVPR*, 2017. 2

[9] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In *NAACL*, 2022. 2

[10] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *arXiv preprint arXiv:2206.01201*, 2022. 2

[11] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, , and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2021. 2

[14] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022. 1, 2

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, 2022. 4